ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

”ΜΑΘΗΜΑΤΙΚΑ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΚΑΙ ΤΩΝ ΑΠΟΦΑΣΕΩΝ”

Έλεγχοι καλής προσαρμογής σε λογιστικά μοντέλα παλινδρόμησης

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Κωνσταντίνα Ν. Στασινοπούλου

Επιβλέπων : Κωνσταντίνος Πετρόπουλος

Πάτρα, Νοέμβριος 2019

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
”ΜΑΘΗΜΑΤΙΚΑ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΚΑΙ ΤΩΝ ΑΠΟΦΑΣΕΩΝ”

Έλεγχοι καλής προσαρμογής σε λογιστικά μοντέλα παλινδρόμησης

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Κωνσταντίνα Ν. Στασινοπούλου

Επιβλέπων : Κωνσταντίνος Πετρόπουλος

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 26η Νοεμβρίου 2019.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Κωνσταντίνος Πετρόπουλος	Βιολέττα Πιπερίγκου	Νικόλαος Τσάντας
Επίκουρος Καθηγητής	Επίκουρη Καθηγήτρια	Καθηγητής

Πάτρα, Νοέμβριος 2019

Πανεπιστήμο Πατρών, Τμήμα Μαθηματικών

Κωνσταντίνα Ν. Στασινοπούλου

Περίληψη

Στόχος της συγκεκριμένης διπλωματικής εργασίας είναι η διερεύνηση των ελέγχων καλής προσαρμογής για το πολυωνυμικό μοντέλο λογιστικής παλινδρόμησης. Αναπτύσσεται μια νέα στρατηγική, η οποία μπορεί να χρησιμοποιηθεί για τον έλεγχο καλής προσαρμογής για οποιοδήποτε τύπο μοντέλου λογιστικής παλινδρόμησης - δυαδικό, πολυωνυμικό, διατακτικό. Επιπλέον, στην περίπτωση που οι έλεγχοι υποδηλώνουν κακή προσαρμογή, είναι εύκολο να εντοπισθούν οι τύποι των ατόμων που δεν μοντελοποιούνται σωστά.

Η παρούσα εργασία αποτελείται από έξι επιμέρους κεφάλαια. Στο Κεφάλαιο 1 γίνεται εισαγωγή της βασικής θεωρίας του δυαδικού λογιστικού μοντέλου. Στο Κεφάλαιο 2 αναπτύσσονται οι έλεγχοι καλής προσαρμογής που έχουν προταθεί για το μοντέλο αυτό, καθώς και περιορισμοί που χρήζουν αυτούς ακατάλληλους. Στο Κεφάλαιο 3 εισάγεται το πολυωνυμικό λογιστικό μοντέλο και ο έλεγχος καλής προσαρμογής που έχει προταθεί για το συγκεκριμένο μοντέλο. Ακόμη, αναλύεται μια νέα στρατηγική, η οποία βασίζεται στην ομαδοποίηση των παρατηρήσεων των επεξηγηματικών μεταβλητών, σύμφωνα με γνωστές μεθόδους συσταδοποίησης για την διεξαγωγή ενός κλασικού X^2 ελέγχου καλής προσαρμογής του Pearson. Στο Κεφάλαιο 4 πραγματοποιείται προσομοίωση στην R, κατά την οποία διερευνάται η ισχύς των προτεινόμενων ελέγχων και του Hosmer-Lemeshow ελέγχου, σύμφωνα με την ικανότητα τους να εντοπίσουν την παράλειψη ενός όρου αλληλεπίδρασης και ενός τετραγωνικού όρου από το πραγματικό λογιστικό μοντέλο. Στο Κεφάλαιο 5 πραγματοποιείται εφαρμογή σε ένα σύνολο δεδομένων, μέσω της οποίας γίνεται σαφής η χρήση των προτεινόμενων ελέγχων για την αξιολόγηση του πολυωνυμικού λογιστικού μοντέλου, ενώ διεξάγονται συμπεράσματα σε συνδυασμό με τα αποτελέσματα των προσομοιώσεων. Τέλος, στο Κεφάλαιο 6 διατυπώνονται συνολικά συμπεράσματα και προβληματισμοί που δημιουργήθηκαν κατά την έρευνα μας πάνω στα λογιστικά μοντέλα παλινδρόμησης.

Λέξεις Κλειδιά

Δυαδική λογιστική παλινδρόμηση, Πολυωνυμική λογιστική παλινδρόμηση, Έλεγχος καλής προσαρμογής, Ομαδοποίηση, Συσταδοποίηση

Abstract

The aim of this dissertation is to investigate the goodness-of-fit tests for the multinomial logistic regression model. A new approach which can be used to assess goodness-of-fit for all three types of logistic regression models - binary, multinomial, ordinal - will be examined and elaborated. Moreover, in case that the tests indicate lack of fit, the types of subjects that are not modelled well can be easily identified through the contigency table.

This thesis consists of six chapters. In Chapter 1, we introduce the basic theory of the binary logistic model. In Chapter 2, we study the goodness-of-fit tests, which have been proposed for the binary logistic model and we mention some limitations that make these tests inproper. In Chapter 3, we introduce the multinomial logistic model and the proposed goodness-of-fit test for this case. Furthermore, a new strategy based on partitioning in the covariate space according to popular clustering methods is analyzed in order to implement a classic Pearson's Chi-Square goodness-of-fit test. In Chapter 4, simulation studies are carried out in R in order to investigate the power of the proposed tests and the Hosmer-Lemeshow test to detect the omission of a quadratic and an interaction term from the true logistic model. In Chapter 5, the application on a real dataset was performed to illustrate the use of goodness-of-fit test for multinomial logistic regression and we drew certain conclusions about the particular application in conjuction with the results of the simulation study. Finally, in Chapter 6, we state overall inferences and concerns that have arisen during our study in logistic regression models.

Keywords

Binary logistic regression, Multinomial logistic regression, Goodness-of-fit test, Grouping, Clustering

στην οικογένεια μου,

στους φίλους μου και

στον Κώστα Α.

Ευχαριστίες

Θα ήθελα αρχικά να ευχαριστήσω τον επιβλέποντα καθηγητή της συγκεκριμένης διπλωματικής εργασίας, Επίκουρο Καθηγητή Κωνσταντίνο Πετρόπουλο για την εμπιστοσύνη που μου έδειξε κατά τη διάρκεια εκπόνησης της συγκεκριμένης εργασίας. Ακόμη, θα ήθελα να ευχαριστήσω την Επίκουρη Καθηγήτρια Βιολέττα Πιπερίγκου και τον Καθηγητή Νικόλαο Τσάντα που δέχτηκαν να αποτελέσουν μέλη της τριμελούς επιτροπής για την αξιολόγηση της παρούσας μεταπτυχιακής διπλωματικής εργασίας.

Ιδιαίτερες ευχαριστίες θα ήθελα να εκφράσω στους γονείς μου Νικόλαο και Αγγελική, καθώς και στον αδερφό μου Βασίλη για την συνεχή τους στήριξη και κατανόηση καθ'όλη την διάρκεια των σπουδών μου.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Λογιστική Παλινδρόμηση

1.1 Εισαγωγή

Το λογιστικό μοντέλο παλινδρόμησης είναι ένα μη γραμμικό μοντέλο και χρησιμοποιείται για την μοντελοποίηση της σχέσης μεταξύ μιας κατηγορικής εξαρτημένης μεταβλητής και μίας ή περισσότερων επεξηγηματικών (ανεξάρτητων) μεταβλητών. Ανήκει στην ευρύτερη κατηγορία των Γενικευμένων Γραμμικών Μοντέλων και χρησιμοποιείται σε περιπτώσεις, στις οποίες επιθυμούμε να προβλέψουμε την απουσία ή παρουσία ενός χαρακτηριστικού, δηλαδή όταν η μεταβλητή απόκρισης είναι δυαδική. Στην περίπτωση μάλιστα που η μεταβλητή απόκρισης είναι κατηγορική με πάνω από δυο επίπεδα, τα οποία δεν έχουν κάποια ιεραρχική διαβάθμιση μεταξύ τους, καταφεύγουμε στην πολυνωνυμική λογιστική παλινδρόμηση [25].

Η λογική της λογιστικής παλινδρόμησης είναι παρόμοια με αυτή της γραμμικής παλινδρόμησης, με την βασική διαφορά ότι η μεταβλητή απόκρισης Y είναι κατηγορική και όχι ποσοτική, Agresti (2007) [1]. Στόχος δεν είναι η πρόβλεψη των πιθανών τιμών της Y που θα παίρνουν τιμές σε κάποιο υποσύνολο του $\mathbb{R}$, αλλά η πρόβλεψη της κατηγορίας που θα ανήκει μια παρατήρηση της αποκρινόμενης Y στο μέλλον, όπως επίσης και η εκτίμηση των πιθανότητων εμφάνισης των κατηγοριών αυτής σε σχέση με τις επεξηγηματικές μεταβλητές. Επομένως, η διαδικασία προβλέψεων είναι διαφορετική από αυτή της γραμμικής παλινδρόμησης. Στην επόμενη ενότητα παρατίθενται οι λόγοι, για τους οποίους ένα γραμμικό μοντέλο παλινδρόμησης δεν είναι ικανό να μοντελοποιήσει την σχέση μεταξύ μιας εξαρτημένης δυαδικής μεταβλητής Y και μιας επεξηγηματικής μεταβλητής X.

1.2 Μοντέλα Παλινδρόμησης στα οποία η Υ είναι δυαδική

Έστω $\boldsymbol{n}$ παρατηρήσεις $(y_1, y_2, .., y_n)$ της μεταβλητής απόκρισης $\boldsymbol{Y} = (Y_1, Y_2, .., Y_n)$. Οι μεταβλητές Y_i είναι δυαδικές , δηλαδή μπορούν να πάρουν τις τιμές 0 ή 1. Συγκεκριμένα

$$Y_i = \begin{cases} 1, & \text{παρουσία χαρακτηριστικού} \quad \text{(επιτυχία)} \\ 0, & \text{απουσία χαρακτηριστικού} \quad \text{(αποτυχία)} \end{cases}$$

δηλαδή, οι τυχαίες μεταβλητές Y_i ακολουθούν την κατανομή Bernoulli με παράμετρο $\boldsymbol{\pi_i}$. Έστω x_i η τιμή της ανεξάρτητης μεταβλητής X_i. Η πιθανότητα επιτυχίας (ή πιθανότητα ύπαρξης του χαρακτηριστικού) είναι $\pi_i = \pi(x_i) = P(Y_i = 1|X_i = x_i)$ και αντίστοιχα $1 - \pi_i = P(Y_i = 0|X_i = x_i)$ η πιθανότητα αποτυχίας αυτού.

Εφόσον $Y_i \sim B(1, \pi_i)$, η μέση τιμή και η διασπορά ισούνται με

$$E(Y_i|x_i) = 1 * \pi_i + 0 * (1 - \pi_i) = \pi_i \tag{1.1}$$

$$Var(Y_i|x_i) = (1 - \pi_i)^2\pi_i + (0 - \pi_i)^2(1 - \pi_i) = \pi_i(1 - \pi_i). \tag{1.2}$$

Έστω ότι προσαρμόζεται το απλό μοντέλο γραμμικής παλινδρόμησης

$$E(Y_i|x_i) = \beta_0 + \beta_1 x_i. \tag{1.3}$$

Εξισώνοντας τις Σχέσεις (1.1) και (1.3) προκύπτει ότι

$$\pi_i = \beta_0 + \beta_1 x_i, \tag{1.4}$$

δηλαδή η αναμενόμενη τιμή $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$ της αποκρινόμενης Y_i είναι ίση με την πιθανότητα π_i, δηλαδή η Y_i παίρνει την τιμή 1. Επιπλέον από τις Σχέσεις (1.1) και (1.2) γίνεται αντιληπτό πως οποιοσδήποτε παράγοντας που επηρεάζει την πιθανότητα, επηρεάζει και συνεπώς αλλάζει την μέση τιμή και διασπορά των παρατηρήσεων.

1.2.1 Προβλήματα που δημιουργούνται όταν η Υ είναι δυαδική

Όταν η μεταβλητή απόκρισης Y είναι δυαδική, το γραμμικό μοντέλο δεν μπορεί να χρησιμοποιηθεί για την πρόβλεψη αυτής, καθώς δημιουργούνται τα εξής προβλήματα [22]:

- **Τα σφάλματα δεν κατανέμονται κανονικά**

  Επειδή η Y_i είναι δυαδική, κάθε σφάλμα $\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$ παίρνει μόνο δυο τιμές

  $$\text{Αν } Y_i = 1 \text{ τότε } \epsilon_i = 1 - (\beta_0 + \beta_1 x_i).$$

  $$\text{Αν } Y_i = 0 \text{ τότε } \epsilon_i = 0 - (\beta_0 + \beta_1 x_i).$$

  Έτσι, δεν ικανοποιείται η υπόθεση της κανονικότητας των σφαλμάτων, όπως ορίζεται στις προϋποθέσεις του κανονικού γραμμικού μοντέλου παλινδρόμησης.

- **Δεν ικανοποιείται η υπόθεση της ομοσκεδαστικότητας των σφαλμάτων**

  Επειδή $\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i) = Y_i - \pi_i$ έπεται ότι $Var(\epsilon_i) = Var(Y_i|x_i)$. Συνεπώς, από την Σχέση (1.2) έπεται ότι

  $$Var(\epsilon_i) = Var(Y_i|x_i) = \pi_i(1 - \pi_i) = (\beta_0 + \beta_1 x_i)(1 - \beta_0 - \beta_1 x_i).$$

  Παρατηρείται ότι για διαφορετικές τιμές x_i, μεταβάλλονται και οι αντίστοιχες διασπορές των σφαλμάτων ϵ_i, γεγονός που απαγορεύει την χρήση της Μεθόδου των Ελαχίστων Τετραγώνων, καθώς προκύπτουν εκτιμητές οι οποίοι δεν είναι βέλτιστοι.

- **Περιορισμός της αναμενόμενης τιμής της Y**

  Από την Σχέση (1.1) και από το γεγονός ότι $0 \leq \pi_i \leq 1$, αφού η π_i είναι πιθανότητα,

  $$0 \leq E(Y_i|x_i) \leq 1.$$

Στο σημείο αυτό αξίζει να σημειωθεί πως τα δυο πρώτα προβλήματα μπορούν εύκολα να επιλυθούν. Ο πρώτος περιορισμός μπορεί να εξαλειφθεί, όταν επιλέγονται δείγματα μεγάλου μεγέθους ($n > 30$). Στην περίπτωση αυτή, η Μέθοδος των Ελαχίστων Τετραγώνων παρέχει εκτιμητές, οι οποίοι είναι ασυμπτωτκά κανονικοί, λόγω του Κεντρικού Οριακού Θεωρήματος, ακόμα και όταν τα σφάλματα δεν κατανέμονται κανονικά. Για την εξάληψη του δεύτερου περιορισμού χρησιμοποιούνται Σταθμισμένα Ελάχιστα Τετράγωνα.

Παρόλ' αυτά, ο τρίτος περιορισμός δεν μπορεί να αντιμετωπισθεί, καθώς καθιστά την συνάρτηση απόκρισης Y μη γραμμική. Στην πραγματικότητα απαιτείται η κατασκευή ενός μοντέλου, στο οποίο η αποκρινόμενη Y να παίρνει τιμές εντός του διαστήματος $(0, 1)$, ενώ ταυτόχρονα η αναμενόμενη τιμή $E(Y|x)$ να παίρνει τιμές σε όλο το $\mathbb{R}$. Επομένως, το γραμμικό μοντέλο παλινδρόμησης δεν είναι ικανό να προβλέψει την Y στην περίπτωση που αυτή είναι δυαδική. Την λύση στο συγκεκριμένο πρόβλημα δίνει η χρήση της *λογιστικής παλινδρόμησης*.

1.2.2 Υποθέσεις Λογιστικής Παλινδρόμησης

Η λογιστική παλινδρόμηση δεν απαιτεί πολλές από τις βασικές προϋποθέσεις των γραμμικών μοντέλων παλινδρόμησης όπως την ύπαρξη γραμμικής σχέσης μεταξύ εξαρτημένης και επεξηγηματικών μεταβλητών, την κανονικότητα και ομοσκεδαστικότητα των σφαλμάτων καθώς και την εξαρτημένη μεταβλητή να μετράται σε διαστημική ή αναλογική κλίμακα. Παρόλ' αυτά θα πρέπει να ισχύουν κάποιες άλλες βασικές προϋποθέσεις [28] :

1. Η δυαδική λογιστική παλινδρόμηση (binary logistic regression) απαιτεί η αποκρινόμενη μεταβλητή Y να είναι δυαδική.

2. Εφόσον εκτιμάται η πιθανότητα $(\pi = P(Y = 1))$ να συμβεί ένα συγκεκριμένο γεγονός, καθίσταται απαραίτητη η σωστή κωδικοποιήση της μεταβλητής απόκρισης Y. Έτσι, το επιθυμητό αποτέλεσμα θα πρέπει να κωδικοποιείται με 1.

3. Οι παρατηρήσεις δεν θα πρέπει να προέρχονται από επαναλαμβάνομενες μετρήσεις ή συζευγμένα δεδομένα (matched data).

4. Στο μοντέλο θα πρέπει να υπάρχει ελάχιστη ή μηδενική πολυσυγγραμμικότητα μεταξύ των επεξηγηματικών μεταβλητών. Αυτό σημαίνει ότι οι ανεξάρτητες μεταβλητές δεν θα πρέπει να είναι υψηλά συσχετισμένες μεταξύ τους.

5. Παρόλο που δεν απαιτείται η ύπαρξη γραμμικής σχέσης μεταξύ της αποκρινόμενης με τις επεξηγηματικές μεταβλητές, είναι απαραίτητο οι επεξηγηματικές μεταβλητές να συνδέονται γραμμικά με τα ln(odds) ενός γεγονότος.

6. Δεν θα πρέπει να υπάρχουν έκτροπες παρατηρήσεις (outliers) στα δεδομένα.

7. Τυπικά απαιτούνται μεγάλα μεγέθη δειγμάτων. Ένας εύκολος κανόνας για τον υπολογισμό του ελάχιστου μεγέθους δείγματος που απαιτείται στην λογιστική παλινδρόμηση Peduzzi et al. (1996) [12] υποδεικνύει ότι αν π είναι η μικρότερη πιθανότητα ενός ατόμου (subject) του πληθυσμού και k είναι το πλήθος των επεξηγηματικών μεταβλητών, τότε το ελάχιστο πλήθος περιστατικών που θα πρέπει να συμπεριληφθεί στην ανάλυση ισούται με $N = \frac{10k}{\pi}$.

1.3 Δυαδικό λογιστικό μοντέλο

Το μοντέλο που χρησιμοποιείται όταν η αποκρινόμενη Y είναι δυαδική καλείται δυαδικό. Το δυαδικό λογιστικό μοντέλο με μια επεξηγηματική μεταβλητή x_i, για λόγους ευκολίας, θα καλείται *απλό λογιστικό μοντέλο* [25] και ορίζεται ως εξής [10]

$$Y_i = E(Y_i|x_i) + \epsilon_i \tag{1.5}$$

$$E(Y_i|x_i) = \pi_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}, \tag{1.6}$$

όπου $Y \sim B(1, \pi_i), \forall i = 1, 2, .., n$

Στην συνέχεια ακολουθεί η απόδειξη του μοντέλου της Σχέσης 1.6. Σύμφωνα με την προηγούμενη ενότητα, η αναμενόμενη συνάρτηση $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$ λαμβάνει τιμές στο $(-\infty, +\infty)$, ενώ ταυτόχρονα θα πρέπει $0 \leq E(Y_i|x_i) = \pi_i \leq 1$. Προφανώς το μοντέλο της Σχέσης (1.4) δεν είναι κατάλληλο, καθώς οι τιμές $\beta_0 + \beta_1 x_i$ δεν βρίσκονται μέσα στο διάστημα $(0, 1)$. Προκειμένου να γίνει κατάλληλο το μοντέλο, πραγματοποιούνται οι εξής μετασχηματισμοί : αρχικά η πιθανότητα $\pi_i \in (0, 1)$ αντικαθίσταται από την σχετική πιθανότητα $odds = \frac{\pi_i}{1-\pi_i} \in [0, \infty)$. Στην συνέχεια λαμβάνεται ο φυσικός λογάριθμος της σχετικής πιθανότητας, δηλαδή $ln(odds) = ln(\frac{\pi_i}{1-\pi_i}) \in (-\infty, +\infty)$. Τελικά, ο μετασχηματισμός

$$\boldsymbol{g(x_i) = ln(odds) = ln(\frac{\pi_i}{1-\pi_i})} \tag{1.7}$$

καλείται *logit μετασχηματισμός* της πιθανότητας π_i.

Το μετασχηματισμένο μοντέλο της Σχέσης (1.3) παίρνει την μορφή

$$g(x_i) = \beta_0 + \beta_1 x_i \tag{1.8}$$

και καλείται *logit μετασχηματισμένο μοντέλο.*

Εξισώνοντας τις Σχέσεις (1.7) και (1.8) αποδεικνύεται τελικά το ζητούμενο.

$$\begin{aligned} ln(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_i \quad &\Leftrightarrow \quad \frac{\pi_i}{1-\pi_i} = e^{\beta_0+\beta_1 x_i} \\ &\Leftrightarrow \quad \pi_i = (1-\pi_i)e^{\beta_0+\beta_1 x_i} \\ &\Leftrightarrow \quad \pi_i = e^{\beta_0+\beta_1 x_i} - \pi_i e^{\beta_0+\beta_1 x_i} \\ &\Leftrightarrow \quad \pi_i(1+e^{\beta_0+\beta_1 x_i}) = e^{\beta_0+\beta_1 x_i} \\ &\Leftrightarrow \quad \pi_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}, \quad \forall i = 1, .., n. \end{aligned}$$

1.3.1 Εκτίμηση παραμέτρων για το απλό λογιστικό μοντέλο

Εφόσον $Y_i \sim B(1, \pi_i)$, η συνάρτηση πυκνότητας πιθανότητας της εξαρτημένης Y_i είναι

$$f_{Y_i}(y_i; \pi_i) = \pi_i^{y_i}(1-\pi_i)^{(1-y_i)} \qquad y_i = 0, 1 \quad \text{και} \quad i = 1, 2, .., n \tag{1.9}$$

και λόγω ανεξαρτησίας των παρατηρήσεων της Y_i, η από κοινού συνάρτηση πυκνότητας πιθανότητας ισούται με

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} f(y_i; \pi_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{(1-y_i)} \tag{1.10}$$

Για την εκτίμηση των παραμέτρων β_0, β_1 του λογιστικού μοντέλου παλινδρόμησης χρησιμοποιείται η *μέθοδος της μέγιστης πιθανοφάνειας (maximum likelihood estimation).*

$$\begin{aligned} lnL(\beta_0, \beta_1) &= ln \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{(1-y_i)} \\ &= \sum_{i=1}^{n} [y_i ln\pi_i + (1-y_i) ln(1-\pi_i)] \\ &= \sum_{i=1}^{n} y_i ln\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} ln(1-\pi_i) \end{aligned} \tag{1.11}$$

Λόγω των Σχέσεων (1.8) και (1.6) με άμεση αντικατάσταση των $ln\dfrac{\pi_i}{1-\pi_i}$ και $1-\pi_i$ προκύπτει

$$\begin{aligned} lnL(\beta_0, \beta_1) &= \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) + \sum_{i=1}^{n} ln(1 - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}) \\ &= \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) + \sum_{i=1}^{n} ln\left[1+e^{\beta_0+\beta_1 x_i}\right]^{-1} \\ &= \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} ln\left(1+e^{\beta_0+\beta_1 x_i}\right). \end{aligned} \tag{1.12}$$

Στο σημείο αυτό αξίζει να σημειωθεί πως η εκτίμηση των β_0, β_1 με την μέθοδο της μέγιστης πιθανοφάνειας, την εύρεση δηλαδή των τιμών b_0, b_1 που μεγιστοποιούν την συνάρτηση (1.12) δεν είναι δυνατή, καθώς δεν υπάρχουν λύσεις κλειστής μορφής για τα b_0, b_1 με την μέθοδο αυτήν. Για την εκτίμηση αυτών θα πρέπει να γίνει χρήση επαναληπτικών αριθμητικών μεθόδων.

Εφόσον βρεθούν οι εκτιμητές παλινδρόμησης b_0, b_1, από την Σχέση (1.6) προκύπτει η προσαρμοσμένη αποκρινόμενη συνάρτηση για την i-παρατήρηση $\forall i = 1, .., n$

$$\hat{\pi_i} = \frac{e^{b_0+b_1x_i}}{1+e^{b_0+b_1x_i}} = \frac{e^{\hat{g}(x_i)}}{1+e^{\hat{g}(x_i)}}, \tag{1.13}$$

όπου

$$\hat{g}(x_i) = ln(\hat{odds}) = ln(\frac{\hat{\pi_i}}{1-\hat{\pi_i}}) = b_o + b_1x_i \tag{1.14}$$

ο logit μετασχηματισμός για το προσαρμοσμένο μοντέλο.

1.3.2 Ερμηνεία του εκτιμώμενου συντελεστή παλινδρόμησης b_1

Ενδιαφέρον παρουσιάζει η ερμηνεία του εκτιμώμενου συντελεστή παλινδρόμησης b_1 για το απλό λογιστικό μοντέλο. Θεωρούμε δυο περιπτώσεις για την επεξηματική μεταβλητή X.

- **Για συνεχή επεξηγηματική μεταβλητή $\boldsymbol{X}$**

Για κάθε μια μονάδα που αυξάνεται η τιμή της X, η εκτιμώμενη σχετική πιθανότητα $\frac{\hat{\pi}}{1-\hat{\pi}}$ πολλαπλασιάζεται με την ποσότητα e^{b_1}. Για να γίνει αυτό αντιληπτό, θεωρούμε τον προσαρμοσμένο logit μετασχηματισμό $\hat{g}(X)$ για τις τυχαίες τιμές x_j και $x_j + 1$, $\forall j$ της επεξηγηματικής X. Παίρνοντας την διαφορά αυτών προκύπτει

$$\hat{g}(x_j+1) - \hat{g}(x_j) = b_0 + b_1 \times (x_j+1) - (b_0 + b_1 \times x_j) = b_1.$$

Λόγω της Σχέσης (1.14) η παραπάνω παίρνει την μορφή

$$ln(\hat{odds}_2) - ln(\hat{odds}_1) = ln\left(\frac{\hat{odds}_2}{\hat{odds}_1}\right) = ln(\hat{OR}) = b_1, \tag{1.15}$$

όπου $\hat{odds}_2, \hat{odds}_1$ αποτελούν εκτιμητές των

$$odds_2 = \frac{P(Y=1|X=x_j+1)}{P(Y=0|X=x_j+1)} \quad \text{και} \quad odds_1 = \frac{P(Y=1|X=x_j)}{P(Y=0|X=x_j)} \quad \text{αντίστοιχα.}$$

Χρησιμοποιώντας την εκθετική συνάρτηση, η Σχέση (1.15) γίνεται,

$$\hat{OR} = \frac{\hat{odds}_2}{\hat{odds}_1} = e^{b_1}. \tag{1.16}$$

Επομένως, η ποσότητα e^{b_1} αποτελεί εκτιμήτρια του λόγου σχετικών πιθανοτήτων OR.

Παρατήρηση 1.3.2.1

Για συνεχή επεξηγηματική μεταβλητή, αν $b_1 > 0$ τότε $e^{b_1} = \hat{OR} > 1$, δηλαδή τα odds ενός γεγονότος αυξάνονται κατά την μοναδιαία αύξηση της ανεξάρτητης X. Κατά αντιστοιχία, αν $b_1 < 0$ τότε $e^{b_1} = \hat{OR} < 1$, δηλαδή τα odds ενός γεγονότος μειώνονται κατά την μοναδιαία αύξηση της ανεξάρτητης X.

- **Για δυαδική (κατηγορική) επεξηγηματική μεταβλητή X**

Έστω η επεξηγηματική X παίρνει τις τιμές 0 και 1. Παίρνοντας την διαφορά των logit μετασχηματισμών για $X = 0$ και $X = 1$ προκύπτει άμεσα

$$\hat{g}(1) - \hat{g}(0) = (b_0 + b_1 \times 1) - (b_0 + b_1 \times 0) = b_1.$$

Ακολουθώντας την παραπάνω διαδικασία για συνεχή επεξηγηματική μεταβλητή, έπεται ομοίως η Σχέση 1.16, όπου $\hat{odds}_2, \hat{odds}_1$ αποτελούν εκτιμητές των

$$odds_2 = \frac{P(Y=1|X=1)}{P(Y=0|X=1)} \quad \text{και} \quad odds_1 = \frac{P(Y=1|X=0)}{P(Y=0|X=0)} \quad \text{αντίστοιχα.}$$

Παρατήρηση 1.3.2.2

Για δυαδική επεξηγηματική μεταβλητή, αν $b_1 > 0$ τότε $e^{b_1} = \hat{OR} > 1$, δηλαδή τα odds στην ομάδα που $X = 1$, είναι μεγαλύτερα σε σχέση με τα odds της ομάδας αυτών που $X = 0$. Κατά αντιστοιχία, αν $b_1 < 0$ τότε $e^{b_1} = \hat{OR} < 1$, δηλαδή τα odds στην ομάδα που $X = 1$, είναι μικρότερα σε σχέση με τα odds της ομάδας αυτών που $X = 0$.

1.4 Πολλαπλή Λογιστική Παλινδρόμηση

Σε αρκετά προβλήματα, η μεταβλητή απόκρισης Y μπορεί να θεωρηθεί ότι επηρεάζεται από περισσότερες από μια επεξηγηματικές μεταβλητές. Έστω $x_1, x_2, ..., x_p$ το πλήθος των p επεξηγηματικών μεταβλητών, οι οποίες μπορεί να είναι κατηγορικές, συνεχείς ή συνδυασμός και των δυο. Στην περίπτωση αυτή το απλό λογιστικό μοντέλο (1.5) - (1.6) γενικεύεται στο πολλαπλό λογιστικό μοντέλο παλινδρόμησης.

Το πολλαπλό λογιστικό μοντέλο για την $i-$παρατήρηση παίρνει την μορφή [10], [23]

$$Y_i = E(Y_i|\boldsymbol{x}_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{1.17}$$

$$E(Y_i|\boldsymbol{x}_i) = \pi(\boldsymbol{x_i}) = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+..+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+..+\beta_p x_{ip}}} = \frac{e^{\boldsymbol{x_i'\beta}}}{1+e^{\boldsymbol{x_i'\beta}}}, \tag{1.18}$$

όπου

$$\underset{n\times 1}{\boldsymbol{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}, \quad \underset{(p+1)\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \underset{n\times 1}{\boldsymbol{\epsilon}} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix} \quad \underset{n\times(p+1)}{\boldsymbol{\mathcal{X}}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}$$

όπου $\boldsymbol{x_i'} = (1, x_{i1}, \ldots, x_{ip})$ το διάνυσμα των επεξηγηματικών μεταβλητών του μοντέλου για την i-οστή παρατήρηση και $\boldsymbol{\mathcal{X}}$ είναι ο πίνακας πληροφορίας του μοντέλου, με την πρώτη στήλη να ισούται με 1 για τον σταθερό όρο β_0 (intercept). Ο αντίστοιχος logit μετασχηματισμός για την i-παρατήρηση του πολλαπλού μοντέλου λογιστικής παλινδρόμησης δίνεται από την εξίσωση

$$g(\boldsymbol{x_i}) = ln(odds) = ln\left(\frac{\pi(\boldsymbol{x_i})}{1-\pi(\boldsymbol{x_i})}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \boldsymbol{x_i'\beta}. \tag{1.19}$$

Η εκτίμηση των παραμέτρων γίνεται και πάλι με την μέθοδο της μέγιστης πιθανοφάνειας. Όπως και στο απλό μοντέλο, έτσι και στο πολλαπλό δεν μπορούν να υπολογισθούν οι εκτιμητές των $\beta_0, \beta_1, \ldots, \beta_p$ σε κλειστή μορφή. Έτσι καθίσταται και πάλι αναγκαία η χρήση επαναληπτικών μεθόδων για την εκτίμηση αυτών.

Ο προσαρμοσμένος logit μετασχηματισμός για την $i-$ παρατήρηση του μοντέλου είναι

$$\hat{g}(\boldsymbol{x_i}) = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip} = \boldsymbol{x_i' b} \tag{1.20}$$

Η ερμηνεία των συντελεστών παλινδρόμησης στο πολλαπλό μοντέλο αποτελεί επέκταση της αντίστοιχης ερμηνείας για το μοντέλο με μια ανεξάρτητη μεταβλητή. Συγκεκριμένα για συνεχείς επεξηγηματικές μεταβλητές, η ποσότητα e^{b_k} με $k = 1, \ldots, p$ είναι ο παράγοντας με τον οποίο πολλαπλασιάζεται η εκτιμώμενη σχετική πιθανότητα $\frac{\hat{\pi}(\boldsymbol{x_k})}{1-\hat{\pi}(\boldsymbol{x_k})}$, για κάθε μονάδα που αυξάνεται η τιμή της X_k, διατηρώντας όλες τις υπόλοιπες συμμεταβλητές σταθερές. Επιπλέον, όταν $b_k > 0$, τα odds ενός γεγονότος αυξάνονται κατά την μοναδιαία αύξηση της

τιμής της ανεξάρτητης X_k και αντίστροφα. Αντίστοιχα, για δυαδικές (κατηγορικές) επεξηγηματικές μεταβλητές, αν $e^{b_k} = \hat{OR} > 1$ σημαίνει ότι τα odds στην ομάδα αυτών που $X_k = 1$, είναι μεγαλύτερα σε σχέση με τα odds της ομάδας αυτών που $X_k = 0$ και αντίστροφα.

1.5 Διωνυμική Λογιστική Παλινδρόμηση

Στην περίπτωση που η αποκρινόμενη μεταβλητή Y_i μετρά τον αριθμό των «επιτυχιών» σε n_i δοκιμές Bernoulli με πιθανότητα επιτυχίας $\pi(\boldsymbol{x_i})$ ίδια σε κάθε δοκιμή και επιπλέον οι δοκιμές είναι ανεξάρτητες μεταξύ τους, τότε η αποκρινόμενη μεταβλητή Y_i ακολουθεί την διωνυμική κατανομή με πιθανότητα επιτυχίας $\pi(\boldsymbol{x_i})$. Σε αντιστοιχία προκύπτει ότι το πλήθος $n_i - Y_i$ των δοκιμών είναι «αποτυχίες» με πιθανότητα αποτυχίας $1 - \pi(\boldsymbol{x_i})$ ίδια σε κάθε δοκιμή, όπου $i = 1, \ldots, n$ το σύνολο των παρατηρήσεων. Συμβολικά $Y_i \sim Bin(n_i, \pi(\boldsymbol{x_i}))$.

Η μέση τιμή και διασπορά του αριθμού των επιτυχιών Y_i σε n_i δοκιμές [15] είναι

$$E(Y_i) = n_i\pi(\boldsymbol{x_i}) = n_i\frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}}$$

$$Var(Y_i) = n_i\pi(\boldsymbol{x_i})(1 - \pi(\boldsymbol{x_i})) = n_i\frac{e^{\boldsymbol{x_i'\beta}}}{(1 + e^{\boldsymbol{x_i'\beta}})^2}.$$

Παρατηρείται και πάλι πως οποιοσδήποτε παράγοντας που θα επηρεάσει την πιθανότητα $\pi(\boldsymbol{x_i})$, επηρεάζει την μέση τιμή και την διασπορά της αποκρινόμενης Y_i.

Παρατήρηση 1.5.1

Στην διωνυμική λογιστική παλινδρόμηση, το ποσοστό επιτυχίας της $i-$οστής παρατήρησης $\frac{Y_i}{n_i}$ χρησιμοποιείται ως μεταβλητή απόκρισης, καθώς το εύρος του ποσοστού $\frac{Y_i}{n_i}$ κυμαίνεται πάντα μεταξύ 0 και 1, ενώ η τιμής της Y_i κυμαίνεται μεταξύ 0 και n_i και είναι διαφορετική για κάθε παρατήρηση.

Παρατήρηση 1.5.2

Η δυαδική λογιστική παλινδρόμηση (binary logistic regression) αποτελεί ειδική περίπτωση της διωνυμικής λογιστικής παλινδρόμησης (binomial logistic regression) για $n_i = 1$.

Κεφάλαιο 2

Έλεγχοι καλής προσαρμογής για το δυαδικό λογιστικό μοντέλο παλινδρόμησης

2.1 Έλεγχοι Pearson και Απόκλισης

Οι έλεγχοι καλής προσαρμογής εξετάζουν πόσο καλά προσαρμόζεται το μοντέλο στα δεδομένα μας. Στην πραγματικότητα, οι έλεγχοι αυτοί συγκρίνουν τις παρατηρούμενες τιμές (observed values) με τις προσαρμοσμένες (fitted values). Η διεξαγωγή των ελέγχων αποτελεί κρίσιμο κομμάτι στη λογιστική παλινδρόμηση, καθώς λανθασμένες αποφάσεις σχετικά με την προσαρμογή του μοντέλου μπορεί να οδηγήσουν σε λανθασμένη ερμηνεία των αποτελεσμάτων της παλινδρόμησης. Προτού γίνει η περιγραφή των βασικών μεθόδων για το μοντέλο αυτό, θα γίνει μια σύντομη αναφορά στην έννοια των «*προτύπων συμμεταβλητών (covariate patterns)*», στην οποία βασίζονται οι δυο έλεγχοι καλής προσαρμογής που θα αναφερθούν παρακάτω.

Ένα πρότυπο συμμεταβλητών (covariate pattern) [7], [10], [21] χρησιμοποιείται για να περιγράψει ένα συγκεριμένο σχηματισμό των τιμών των συμμεταβλητών του μοντέλου. Για παράδειγμα, εάν στο μοντέλο υπάρχουν μόνο δυο κατηγορικές επεξηγηματικές μεταβλητές, έστω το φύλο και την οικογενειακή κατάσταση, κάθε μια με δυο επίπεδα, τότε δημιουργούνται μόνο τέσσερις δυνατοί συνδυασμοί των τιμών των συμμεταβλητών και επομένως τέσσερα δυνατά πρότυπα συμμεταβλητών. Όταν όμως στο μοντέλο υπάρχουν συνεχείς συμ-

μεταβλητές, τότε κάθε παρατήρηση πιθανώς να οδηγεί στη δημιουργία ενός διαφορετικού προτύπου. Αυτό θα δούμε παρακάτω πως αποτελεί εμπόδιο για την εφαρμογή των ελέγχων Pearson και Απόκλισης, στην περίπτωση που υπάρχει τουλάχιστον μια συνεχής επεξηγηματική μεταβλητή στο μοντέλο.

Έστω $\boldsymbol{\mathcal{X}}$ ο πίνακας πληροφορίας του μοντέλου και έστω J το πλήθος των προτύπων συμμεταβλητών για το $\boldsymbol{\mathcal{X}}$. Εαν κάποια «άτομα (subjects)» στα δεδομένα έχουν μεταξύ τους τις ίδιες τιμές των συμμεταβλητών, τότε ισχύει $J < n$. Συμβολίζουμε με m_j το πλήθος των ατόμων με $x = x_j$, $j = 1, \ldots, J$. Έπεται ότι $\sum_{j=1}^{J} m_j = n$. Έστω ακόμη y_j να είναι το πλήθος των επιτυχιών $(Y = 1)$ μεταξύ των m_j ατόμων για το συγκεκριμένο πρότυπο. Άμεσα έπεται ότι $\sum_{j=1}^{J} y_j = n_1$ είναι το συνολικό πλήθος επιτυχιών για όλα τα άτομα του δείγματος και επομένως $n_1 + n_0 = n$, όπου n_0 το αντίστοιχο πλήθος των αποτυχιών $(Y = 0)$. Στην λογιστική παλινδρόμηση, οι προσαρμοσμένες τιμές υπολογίζονται για κάθε πρότυπο συμμεταβλητών και εξαρτώνται από την εκτιμώμενη πιθανότητα του κάθε προτύπου.

Συνεπώς για το j-πρότυπο συμμεταβλητών οι προσαρμοσμένες τιμές [17] είναι ,

$$\hat{y_j} = m_j \hat{\pi}(x_j) = m_j \left[\frac{e^{\hat{g}(x_j)}}{1 + e^{\hat{g}(x_j)}} \right], \tag{2.1}$$

όπου $\hat{g}(x_j) = b_0 + b_1 x_{j1} + \ldots + b_p x_{jp}$ ο προσαρμοσμένος logit μετασχηματισμός.

Αξίζει να τονισθεί πως το πλήθος J των προτύπων δεν επηρεάζει την κατασκευή του μοντέλου. Οι βαθμοί ελευθερίας για την διεξαγωγή των ελέγχων βασίζονται στο πλήθος των παραμέτρων των μοντέλων προς σύγκριση και μόνο. Το πλήθος των προτύπων αποτελεί ζήτημα όταν πρόκειται για τον έλεγχο καλής προσαρμογής του μοντέλου.

Η συνάρτηση πιθανοφάνειας και αντίστοιχα ο φυσικός της λογάριθμος με βάση τα πρότυπα συμμεταβλητών δίνονται από τις παρακάτω σχέσεις [21]

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{J} \binom{m_j}{y_j} (\pi(x_j))^{y_j} ((1 - \pi(x_j))^{m_j - y_j} \tag{2.2}$$

$$lnL(\boldsymbol{\beta}) = \sum_{j=1}^{J} \left[ln\binom{m_j}{y_j} + y_j ln\pi(x_j) + (m_j - y_j) ln(1 - \pi(x_j)) \right]. \tag{2.3}$$

Ο ζητούμενος έλεγχος καλής προσαρμογής είναι

H_0 : το μοντέλο της λογιστικής παλινδρόμησης παρουσιάζει ικανοποιητική προσαρμογή έναντι της

H_1 : το μοντέλο της λογιστικής παλινδρόμησης δεν παρουσιάζει ικανοποιητική προσαρμογή

Παρακάτω αναπτύσσονται οι τρεις πιο διαδεδομένοι έλεγχοι καλής προσαρμογής που έχουν προταθεί στη βιβλιογραφία για το δυαδικό λογιστικό μοντέλο παλινδρόμησης.

2.1.1 Ο έλεγχος X^2 του Pearson

Για το j-πρότυπο συμμεταβλητών, το κατάλοιπο του Pearson (Pearson's residual) [17] είναι

$$r(y_j, \hat{\pi}(x_j)) = \frac{(y_j - m_j\hat{\pi}(x_j))}{\sqrt{m_j\hat{\pi}(x_j)(1-\hat{\pi}(x_j))}}. \tag{2.4}$$

Η ελεγχοσυνάρτηση Pearson ορίζεται ως εξής

$$X^2 = \sum_{j=1}^{J} r(y_j, \hat{\pi}(x_j))^2. \tag{2.5}$$

Το κατάλοιπο του Pearson υπολογίζει την διαφορά μεταξύ των παρατηρούμενων τιμών της εξαρτημένης Y με τις αντίστοιχες εκτιμώμενες πιθανοτήτες κάθε προτύπου και στην συνέχεια το X^2 της Σχέσης (2.5) αθροίζει για όλα τα J πρότυπα συμμεταβλητών. Μικρή τιμή του X^2 υποδηλώνει πως το μοντέλο προσαρμόζεται καλά στα δεδομένα, ενώ μεγάλη τιμή αποτελεί ένδειξη απόκλισης από το σωστό μοντέλο.

Η ελεγχοσυνάρτηση X^2 μπορεί να προκύψει και ως αποτέλεσμα από έναν πίνακα συνάφειας $J \times 2$, όπως φαίνεται στον Πίνακα 2.1. Έστω O_{1j} και E_{1j} η παρατηρούμενη και εκτιμώμενη αναμενόμενη τιμή με $Y = 1$ για το j-πρότυπο. Κατά αντιστοιχία O_{0j} και E_{0j} είναι η παρατηρούμενη και εκτιμώμενη αναμενόμενη τιμή με $Y = 0$ για το j-πρότυπο. Οι εκτιμώμενες αναμενόμενες τιμές υπό την μηδενική υπόθεση ισούνται με $m_j\hat{\pi}(x_j)$ για $Y = 1$ και αντίστοιχα $m_j(1 - \hat{\pi}(x_j))$ για $Y = 0$.

Τότε, η ελεγχοσυνάρτηση X^2 του Pearson υπολογίζεται ως εξής

$$X^2 = \sum_{i=0}^{1}\sum_{j=0}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \tag{2.6}$$

Πίνακας 2.1: *Πίνακας συνάφειας για τον έλεγχο καλής προσαρμογής για* $Y = 0, 1$

	$Y=0$		$Y=1$		
Πρότυπο Συμμεταβλητών	Παρ.	Εκτ.	Παρ.	Εκτ.	Σύνολο
1	O_{01}	E_{01}	O_{11}	E_{11}	n_1
2	O_{02}	E_{02}	O_{12}	E_{12}	n_2
⋮	⋮		⋮		⋮
J	O_{0J}	E_{0J}	O_{1J}	E_{1J}	n_J

Υπό την μηδενική υπόθεση H_0 ότι το μοντέλο προσαρμόζεται καλά στα δεδομένα, η ελεγχοσυνάρτηση X^2 ακολουθεί ασυμπτωτικά την χ^2 κατανομή με $J-(p+1)$ βαθμούς ελευθερίας, όπου p είναι το πλήθος των παραμέτρων του μοντέλου (χωρίς τον β_0).

2.1.2 Έλεγχος Απόκλισης

Η Απόκλιση (Deviance) παίζει σημαντικό ρόλο στην αξιολόγηση της προσαρμογής του μοντέλου στα δεδομένα, καθώς μετράει την συνολική μεταβλητότητα που μένει ανεξήγητη από το προσαρμοσμένο μοντέλο. Η Απόκλιση D για την λογιστική παλινδρόμηση είναι ισοδύναμη του SSE για την γραμμική παλινδρόμηση. Η απόκλιση ορίζεται ως

$$D = -2ln\left[\frac{\text{πιθανοφάνεια προσαρμοσμένου μοντέλου}}{\text{πιθανοφάνεια κορεσμένου μοντέλου}}\right] = -2\left(\frac{L_{\boldsymbol{f}}}{L_{\boldsymbol{s}}}\right). \tag{2.7}$$

Το κορεσμένο (saturated) μοντέλο είναι το μοντέλο που περιέχει τόσες ανεξάρτητες μεταβλητές όσες και ο αριθμός των παρατηρήσεων. Είναι με άλλα λόγια το μοντέλο, το οποίο προσαρμόζεται τέλεια στα δεδομένα. Έστω $\hat{\pi}_{\boldsymbol{s}}(x_j)$ να είναι ο ΕΜΠ για το κορεσμένο μοντέλο $\hat{L}_{\boldsymbol{s}}$ και αντίστοιχα $\hat{\pi}(x_j)$ ο ΕΜΠ για το προσαρμοσμένο μοντέλο $\hat{L}_{\boldsymbol{f}}$. Οι μεγιστοποιημένοι λογάριθμοι της ΕΜΠ για τα δυο μοντέλα μέσω της Σχέσης (2.3) είναι αντίστοιχα

$$ln\hat{L}_{\boldsymbol{s}} = \sum_{j=1}^{J}\left[ln\binom{m_j}{y_j} + y_j ln\hat{\pi}_{\boldsymbol{s}}(x_j) + (m_j - y_j)ln(1-\hat{\pi}_{\boldsymbol{s}}(x_j))\right] \tag{2.8}$$

$$ln\hat{L}_{\boldsymbol{f}} = \sum_{j=1}^{J}\left[ln\binom{m_j}{y_j} + y_j ln\hat{\pi}(x_j) + (m_j - y_j)ln(1-\hat{\pi}(x_j))\right] \tag{2.9}$$

όπου $\hat{\pi}_{\boldsymbol{s}}(x_j) = \dfrac{y_j}{m_j}$ και $\hat{\pi}(x_j) = \dfrac{\hat{y}_j}{m_j}$ αντίστοιχα από την Σχέση (2.1).

Τελικά, η ελεγχοσυνάρτηση Απόκλισης μέσω της Σχέσης (2.7) παίρνει την μορφή

$$D = -2\sum_{j=1}^{J}\left[y_j ln\left(\frac{\hat{\pi}(x_j)}{\hat{\pi}_{\boldsymbol{s}}(x_j)}\right) + (m_j - y_j) ln\left(\frac{1-\hat{\pi}(x_j)}{1-\hat{\pi}_{\boldsymbol{s}}(x_j)}\right)\right] \tag{2.10}$$

$$= -2\sum_{j=1}^{J}\left[y_j ln\left(\frac{\hat{y}_j}{y_j}\right) + (m_j - y_j) ln\left(\frac{m_j-\hat{y}_j}{m_j-y_j}\right)\right]. \tag{2.11}$$

και συγκρίνει τις παρατηρούμενες τιμές y_j με τις προσαρμοσμένες τιμές $\hat{y}_j$ για το j-πρότυπο. Στην πραγματικότητα, το D μετρά κατά πόσο αποκλίνει το υπό μελέτη μοντέλο από το κορεσμένο (τέλειο) μοντέλο. Υπό την μηδενική υπόθεση H_0 ότι το μοντέλο προσαρμόζεται καλά στα δεδομένα, η ελεγχοσυνάρτηση Απόκλισης ακολουθεί ασυμπτωτικά την χ^2 κατανομή με $J-(p+1)$ βαθμούς ελευθερίας.

Διαφορετικά, η ελεγχοσυνάρτηση Απόκλισης μπορεί να υπολογισθεί με βάση τα κατάλοιπα Απόκλισης (Deviance residual). Το κατάλοιπο Απόκλισης για το j-πρότυπο συμμεταβλητών ορίζεται ως εξής [10] :

- Για $0 < y_j < m_j$

$$d(y_j, \hat{\pi}(x_j)) = \pm\left\{2\left[y_j ln\left(\frac{y_j}{m_j\hat{\pi}(x_j)}\right) + (m_j - y_j) ln\left(\frac{m_j-y_j}{m_j(1-\hat{\pi}(x_j))}\right)\right]\right\}^{1/2}, \tag{2.12}$$

όπου το πρόσημο της $d(y_{j1}, \hat{\pi}(x_j))$ είναι ίδιο με το πρόσημο της ποσότητας $(y_j - m_j\hat{\pi}(x_j))$.

- Για πρότυπα συμμεταβλητών με $y_j = 0$, το κατάλοιπο Απόκλισης ορίζεται ως

$$d(y_j, \hat{\pi}(x_j)) = -\sqrt{2m_j|ln(1-\hat{\pi}(x_j))|}, \tag{2.13}$$

- ενώ για πρότυπα με $y_j = m_j$, το κατάλοιπο Απόκλισης ισούται με

$$d(y_j, \hat{\pi}(x_j)) = \sqrt{2m_j|ln\hat{\pi}(x_j)|}. \tag{2.14}$$

Η ελεγχοσυνάρτηση Απόκλισης, βασιζόμενη στα κατάλοιπα Απόκλισης προκύπτει ως

$$D = \sum_{j=1}^{J}(d(y_j, \hat{\pi}(x_j)))^2. \tag{2.15}$$

Τέλος, η ελεγχοσυνάρτηση D μπορεί να υπολογισθεί και από τον Πίνακα Συνάφειας 2.1

$$D = 2\sum_{i=0}^{1}\sum_{j=1}^{J} O_{ij}\, ln\frac{O_{ij}}{E_{ij}}. \tag{2.16}$$

Ισοδύναμα με πριν $D \sim \chi^2_{J-(p+1)}$.

Παρόλο που οι ποσότητες X^2 και D διατίθενται σε πολλά στατιστικά πακέτα, οι συγκεκριμένοι έλεγχοι πολλές φορές όχι μόνο δεν παρέχουν πληροφορία για την προσαρμογή του μοντέλου στα δεδομένα, αλλά επίσης η χρήση τους μπορεί να οδηγήσει ακόμα και σε λανθασμένες αποφάσεις [7], [21]. Συγκεκριμένα στις περιπτώσεις όπου :

I) $J \approx n$, δηλαδή όταν το πλήθος των προτύπων είναι σχεδόν ίδιο με το πλήθος των παρατηρήσεων (δηλαδή $m_j = 1$) και η μεταβλητή απόκρισης y_j είναι δυαδική, τότε τόσο ο έλεγχος του Pearson όσο και ο έλεγχος Απόκλισης δεν παρέχουν πληροφορία για την προσαρμογή του μοντέλου, καθώς και οι δυο ελεγχοσυναρτήσεις βασίζονται μόνο στις εκτιμώμενες πιθανότητες και στο μέγεθος του δείγματος και όχι στην σύγκριση των παρατηρούμενων y_j και των εκτιμώμενων $\hat{y}_j$ τιμών, όπως φαίνεται αναλυτικά παρακάτω.

- Για τον έλεγχο του Pearson στην περίπτωση που $J = n$ και $y_j = 0, 1$, η ελεγχοσυνάρτηση X^2 μέσω των Σχέσεων (2.4) και (2.5) είναι η ακόλουθη

$$X^2 = \sum_{j=1}^{n} \frac{(y_j - \hat{\pi}(x_j))^2}{\hat{\pi}(x_j)(1 - \hat{\pi}(x_j))}. \tag{2.17}$$

- Γα τον έλεγχο Απόκλισης, η ελεγχοσυνάρτηση D για την παραπάνω περίπτωση προκύπτει ακολουθώντας τα βήματα :

Στο απλό λογιστικό μοντέλο, ο μεγιστοποιημένος λογάριθμος του ΕΜΠ (1.11) είναι

$$ln\hat{L}_{\boldsymbol{f}} = \sum_{j=1}^{n} [y_j ln\hat{\pi}(x_j) + (1 - y_j)ln(1 - \hat{\pi}(x_j))],$$

ενώ στο κορεσμένο μοντέλο με $\hat{\pi}_{\boldsymbol{s}}(x_j) = y_j$ ισχύει

$$ln\hat{L}_{\boldsymbol{s}} = \sum_{j=1}^{n} [y_j ln(y_j) + (1 - y_{j1})ln(1 - y_j)] = 0.$$

Επειδή η y_j είναι δυαδική, οι ποσότητες $y_j ln(y_j)$ και $(1 - y_j)ln(1 - y_j)$ μηδενίζονται πάντα. Συνεπώς, η ελεγχοσυνάρτηση Απόκλισης ισούται

$$D = -2\sum_{j=1}^{n} [y_j ln\hat{\pi}(x_j) + (1 - y_j)ln(1 - \hat{\pi}(x_j))],$$

ενώ μετά από πράξεις το παραπάνω γράφεται τελικά (Collett (1991) [3])

$$D = -2\sum_{j=1}^{n}[\hat{\pi}(x_j)logit(\hat{\pi}(x_j)) + ln(1-\hat{\pi}(x_j))]. \tag{2.18}$$

Παρατηρούμε πράγματι πως οι ελεγχοσυναρτήσεις X^2 και D μέσω των Σχέσεων (2.17) και (2.18) δεν συγκρίνουν τις παρατηρούμενες τιμές με τις αντίστοιχες προσαρμοσμένες, καθώς εξαρτώνται αποκλειστικά από τις εκτιμώμενες πιθανότητες $\hat{\pi}(x_j)$. Επομένως, οι παραπάνω έλεγχοι όταν η αποκρινόμενη μεταβλητή είναι δυαδική και επιπλέον $J = n$ δεν ισχύουν ως έλεγχοι καλής προσαρμογής.

II) Η ασυμπτωτική χ^2 κατανομή τους υπό την μηδενική υπόθεση ισχύει στην περίπτωση όπου το πλήθος των παρατηρήσεων m_j σε κάθε πρότυπο συμμεταβλητών είναι μεγάλο, έτσι ώστε και οι εκτιμώμενες αναμενόμενες τιμές να είναι "μεγάλες" σε κάθε κελί του πίνακα συνάφειας, δηλαδή $m_j\hat{\pi}(x_j) > 5$. Βέβαια κάτι τέτοιο συνήθως δεν υφίσταται, όταν στο μοντέλο υπάρχει μεγάλο πλήθος κατηγορικών επεξηγηματικών μεταβλητών ή τουλάχιστον μια συνεχής επεξηγηματική μεταβλητή.

Γενικά, όταν στο μοντέλο περιέχονται πολλές κατηγορικές συμμεταβλητές ή τουλάχιστον μια συνεχής επεξηγηματική μεταβλητή, τότε κάθε παρατήρηση οδηγεί στην δημιουργία ενός διαφορετικού προτύπου συμμεταβλητών και άρα $J \approx n$. Αυτό έχει ως αποτέλεσμα, η ασυμπτωτική χ^2 κατανομή των ελέγχων Pearson (2.6) και Απόκλισης (2.16) υπό την μηδενική υπόθεση ότι το μοντέλο έχει καλή προσαρμογή να μην ισχύει, όταν οι αναμενόμενες τιμές στα κελιά του Πίνακα Συνάφειας 2.1 είναι πολύ μικρές [14]).

2.2 Έλεγχος Hosmer και Lemeshow $\hat{C}_g$

Οι Hosmer και Lemeshow (HL) το 1980 ανέπτυξαν επτά διαφορετικές μεθόδους για τον έλεγχο καλής προσαρμογής του λογιστικού μοντέλου παλινδρόμησης [10]. Θα αναφερθούμε στην πιο διαδεδομένη μέθοδο, η οποία βασίζεται αποκλειστικά στην ομαδοποίηση των εκτιμώμενων πιθανοτήτων επιτυχίας του προσαρμοσμένου μοντέλου.

Σε αυτήν την μέθοδο, τα άτομα (subjects) ταξινομούνται σε g ομάδες, όπου σε κάθε ομάδα περιέχονται n/g άτομα. Η μέθοδος αυτή για $g = 10$ είναι γνωστή στην ξένη βιβλιογραφία ως *«deciles of risk»*. Σύμφωνα με αυτή, η πρώτη ομάδα n_1 περιέχει τα $n/10$ άτομα με τις μικρότερες εκτιμώμενες πιθανότητες επιτυχίας που έχουν προκύψει από το προσαρμοσμένο

υπό μελέτη μοντέλο. Η δεύτερη ομάδα n_2 περιέχει τα $n/10$ άτομα με τις αμέσως μικρότερες εκτιμώμενες πιθανότητες. Ισοδύναμα, η δέκατη ομάδα n_{10} περιέχει τα $n/10$ άτομα με τις μεγαλύτερες εκτιμώμενες πιθανότητες. Θεωρούμε O_{kj} να είναι το πλήθος των παρατηρήσεων με $Y = k$ στην $j-$ομάδα και έστω E_{kj} ο αναμενόμενος αριθμός παρατηρήσεων με $Y = k$ για την $j-$ομάδα. Έστω ακόμη n_j το πλήθος των παρατηρήσεων στην $j-$ομάδα με $j = 1, \ldots, g$. Για το δυαδικό λογιστικό μοντέλο με $k = 0, 1$ έπεται ότι :

$$O_{1j} = \sum_{i=1}^{n_j} Y_i \qquad \text{και} \qquad E_{1j} = \sum_{i=1}^{n_j} \hat{\pi}(x_i) \tag{2.19}$$

$$O_{0j} = \sum_{i=1}^{n_j} (1 - Y_i) \qquad \text{και} \qquad E_{0j} = \sum_{i=1}^{n_j} (1 - \hat{\pi}(x_i)). \tag{2.20}$$

Πίνακας 2.2: *Πίνακας συνάφειας για τον έλεγχο καλής προσαρμογής για* $Y = 0, 1$

	$Y = 0$		$Y = 1$		
Ομάδα	Παρ.	Εκτ.	Παρ.	Εκτ.	Σύνολο
1	O_{01}	E_{01}	O_{11}	E_{11}	n_1
2	O_{02}	E_{02}	O_{12}	E_{12}	n_2
⋮	⋮		⋮		⋮
g	O_{0g}	E_{0g}	O_{1g}	E_{1g}	n_g

Η ελεγχοσυνάρτηση των HL υπολογίζεται, εφαρμόζοντας την στατιστική συνάρτηση του Pearson στον $g \times 2$ πίνακα συνάφειας των παρατηρούμενων και αναμενόμενων συχνοτήτων, που απεικονίζεται μέσω του Πίνακα 2.2

$$\hat{C}_g = \sum_{j=1}^{g} \sum_{k=0}^{1} \frac{(O_{kj} - E_{kj})^2}{E_{kj}}. \tag{2.21}$$

Υπό την μηδενική υπόθεση ότι το μοντέλο παρουσιάζει ικανοποιητική προσαρμογή στα δεδομένα, η ελεγχοσυνάρτηση $\hat{C}_g$ ακολουθεί ασυμπτωτικά την χ^2 κατανομή με $g - 2$ βαθμούς ελευθερίας [10].

Ο έλεγχος HL είναι ευρέως διαδεδομένος εξαιτίας των επιθυμητών του ιδιοτήτων. Αρχικά, ο έλεγχος δεν απαιτεί το πλήθος των προτύπων συμμεταβλητών να είναι μικρότερο από το συνολικό πλήθος των ατόμων, δηλαδή $J < n$, όπως γίνεται με τους ελέγχους Pearson

και Απόκλισης,. Όταν στο μοντέλο υπάρχουν συνεχείς συμμεταβλητές, τότε οι έλεγχοι των Pearson και Απόκλισης για τον έλεγχο καλής προσαρμογής του λογιστικού μοντέλου παλινδρόμησης δεν μπορούν να εφαρμοσθούν. Αυτό συμβαίνει διότι κάθε άτομο οδηγεί στην δημιουργία ενός διαφορετικού προτύπου $(m_j = 1)$ με αποτέλεσμα να μην ισχύει η ασυμπτωτική χ^2 κατανομή των ελέγχων, όταν οι συχνότητες στα κελιά του πίνακα συνάφειας είναι πολύ μικρές. Στην περίπτωση αυτή μπορεί να χρησιμοποιηθεί ο έλεγχος των Hosmer και Lemeshow. Επιπλέον, η δημιουργία των ομάδων για την διεξαγωγή του ελέγχου βασίζεται εξ' ολοκλήρου στις εκτιμώμενες πιθανότητες των ατόμων και όχι στα πρότυπα συμμεταβλητών, πράγμα που σημαίνει πως μπορεί να χρησιμοποιηθεί για οποιουσδήποτε τύπους επεξηγηματικών μεταβλητών (συνεχείς,κατηγορικές, κ.α.).

Παρόλ΄αυτά ο συγκεκριμένος έλεγχος παρουσιάζει οριμένα μειονεκτήματα. Αρχικά, ενώ μας πληροφορεί για την συνολική προσαρμογή του μοντέλου, αδυνατεί να ορίσει το βαθμό προσαρμογής αυτού στα δεδομένα. Όταν υποδεικνύει έλλειψη καλής προσαρμογής, αδυνατεί να ορίσει που αυτή οφείλεται. Με άλλα λόγια, ο έλεγχος αδυνατεί να εντοπίσει τα άτομα που δεν μοντελοποιούνται σωστά. Επιπλέον, επηρεάζεται σημαντικά τόσο από το μέγεθος του δείγματος όσο και από το πλήθος των ομάδων που δημιουργούνται . Όσο μεγαλώνει το δείγμα, ο έλεγχος HL σχεδόν πάντα απορρίπτει την μηδενική υπόθεση, γεγονός που υποδεικνύει πάντα κακή προσαρμογή του μοντέλου στα δεδομένα. Αλλά και για μικρότερα μεγέθη μπορεί να αποδειχθεί προβληματικός, αφού ως ένας X^2 έλεγχος απαιτεί υψηλό μέγεθος δείγματος, έτσι ώστε οι αναμενόμενες συχνότητες σε όλα τα κελιά του πίνακα συνάφειας να είναι μεγαλύτερες του 1 και επιπλέον λιγότερο από το 20% των κελιών να έχουν αναμενόμενες συχνότητες μικρότερες του 5 [23]. Ακόμη, έχει παρατηρηθεί πως η επιλογή του αριθμού των ομάδων μπορεί να οδηγήσει κάθε φορά σε διαφορετική απόφαση [19].

Κεφάλαιο 3

Πολυωνυμική Λογιστική Παλινδρόμηση

Στα προηγούμενα κεφάλαια εστιάσαμε στην χρήση του λογιστικού μοντέλου παλινδρόμησης, όταν η αποκρινόμενη μεταβλητή Y είναι δυαδική. Το μοντέλο αυτό εύκολα μπορεί να τροποποιηθεί για την περίπτωση που η μεταβλητή απόκρισης είναι κατηγορική με πάνω από δυο επίπεδα. Τα επίπεδα της αποκρινόμενης Y δεν έχουν κάποια ιεραρχική διαβάθμιση π.χ καλό - πολύ καλό - άριστο. Στην περίπτωση αυτή, το κατάλληλο μοντέλο λογιστικής παλινδρόμησης θα ήταν η διατακτική λογιστική παλινδρόμηση (ordinal logistic regression) (δεν θα αναφερθούμε περαιτέρω στην παρούσα διπλωματική). Για να γίνει κατανοητή η ιδέα του μοντέλου, θεωρούμε μια μελέτη σχετικά με την επιλογή ενός προγράμματος υγείας μεταξύ τεσσάρων προγραμμάτων "Α", "Β", "Γ", "Δ" που προσφέρονται στους υπαλλήλους μιας εταιρείας. Πιθανές επεξηγηματικές μεταβλητές μπορεί να είναι το φύλο, η ηλικία, το εισόδημα και άλλα. Στόχος είναι η εκτίμηση της πιθανότητας επιλογής ενός εκ των τεσσάρων προγραμμάτων. Η *πολυωνυμική λογιστική παλινδρόμηση (multinomial or polytomous or polychotomous logistic regression)* αποτελεί επέκταση της δυαδικής λογιστικής παλινδρόμησης και χρησιμοποιείται σε αυτές τις περιπτώσεις. Για την ανάπτυξη του μοντέλου αρκεί να θεωρήσουμε πως η μεταβλητή απόκρισης Y_i είναι κατηγορική με $c > 2$ επίπεδα, όπου $Y_i = 0, 1, \ldots, c-1$ και $\pi_{ij} = P(Y_i = j|\boldsymbol{x_i})$ είναι η πιθανότητα η i-παρατήρηση να ανήκει στην j-κατηγορία με $\sum_{j=0}^{c-1} \pi_{ij} = 1$. Επιπλέον μία από τις κατηγορίες της αποκρινόμενης μεταβλητής επιλέγεται ως βασική ή προς σύγκριση ή αλλιώς κατηγορία αναφοράς (baseline or reference category). Εφόσον τα επίπεδα της αποκρινόμενης μεταβλητής δεν υπακούουν σε κάποια διάταξη, οποιαδήποτε από τις c-κατηγορίες μπορεί να επιλεχθεί ως κατηγορία αναφοράς. Γενικά, όταν η αποκρινόμενη Y_i έχει c επίπεδα, τότε το μοντέλο αποτελείται από

$c-1$ εξισώσεις και συγκεκριμένα $c-1$ δυαδικά λογιστικά μοντέλα παλινδρόμησης. Κάθε μοντέλο εκφράζει την επιρροή που έχουν οι επεξηγηματικές μεταβλητές στην πιθανότητα επιτυχίας για την συγκεκριμένη κατηγορία σε σχέση με την κατηγορία αναφοράς. Επιπλέον, κάθε μοντέλο έχει διαφορετικούς συντελεστές παλινδρόμησης $\boldsymbol{\beta_j} = (\beta_{j0}, \ldots, \beta_{jp})'$, καθώς οι επεξηγηματικές μεταβλητές επηρεάζουν με διαφορετικό τρόπο κάθε κατηγορία j.

Έστω $Y_i = 0$ να είναι η κατηγορία αναφοράς και $\boldsymbol{x_i'} = (1, x_{i1}, \ldots, x_{ip})$ το διάνυσμα των επεξηγηματικών μεταβλητών του μοντέλου για την i-οστή παρατήρηση. Ο logit μετασχηματισμός για την i-παρατήρηση της $j-$κατηγορίας υπολογίζεται ως εξής (Fagerland et al., (2008) [4] ; Hosmer et al., (2013) [10])

$$g_j(\boldsymbol{x_i}) = ln\left(\frac{\pi_{ij}}{\pi_{i0}}\right) = \beta_{j0} + \beta_{j1}x_{i1} + \ldots + \beta_{jp}x_{ip} = \boldsymbol{x_i'\beta_j}, \quad i = 1, \ldots, n\,; j = 0, \ldots, c-1. \tag{3.1}$$

Οι δεσμευμένες πιθανότητες π_{ij} με $j \neq 0$ υπολογίζονται από τον τύπο

$$\pi_{ij} = P(Y_i = j|\boldsymbol{x_i}) = \frac{e^{\boldsymbol{x_i'\beta_j}}}{1 + \sum_{k=1}^{c-1} e^{\boldsymbol{x_i'\beta_k}}}. \tag{3.2}$$

Η πιθανότητα π_{i0} για την κατηγορία αναφοράς με $j = 0$ είναι αντίστοιχα

$$\pi_{i0} = P(Y_i = 0|\boldsymbol{x_i}) = \frac{1}{1 + \sum_{k=1}^{c-1} e^{\boldsymbol{x_i'\beta_k}}}. \tag{3.3}$$

Η συνάρτηση πιθανοφάνειας μπορεί να γραφεί

$$L(\boldsymbol{\beta}) \simeq \prod_{i=1}^{n}\prod_{j=0}^{c-1}(\pi_{ij})^{Y_i}. \tag{3.4}$$

Για την εύρεση των εκτιμητών $\boldsymbol{b}$ ακολουθείται και πάλι η ίδια διαδικασία που αναφέρθηκε στο Κεφάλαιο 1 για την εκτίμηση των παραμέτρων του απλού λογιστικού μοντέλου παλινδρόμησης (για λεπτομέρειες Hosmer and Lemeshow (2013) [10]).

3.1 Έλεγχος καλής προσαρμογής με βάση την ομαδοποιήση των εκτιμώμενων πιθανοτήτων

Οι περισσότεροι έλεγχοι καλής προσαρμογής για την λογιστική παλινδρόμηση έχουν σχεδιασθεί για την περίπτωση που η μεταβλητή απόκρισης είναι δυαδική. Ελάχιστοι έλεγ-

Πίνακας 3.1: *Πίνακας συνάφειας για τον έλεγχο καλής προσαρμογής για* $Y = 0, 1, \ldots, c-1$

	$Y=0$		$Y=1$		…	$Y=c-1$	
Ομάδα	Παρ.	Εκτ.	Παρ.	Εκτ.	…	Παρ.	Εκτ.
1	O_{10}	E_{10}	O_{11}	E_{11}	…	$O_{1,c-1}$	$E_{1,c-1}$
2	O_{20}	E_{20}	O_{21}	E_{21}	…	$O_{2,c-1}$	$E_{2,c-1}$
⋮					⋱		
g	O_{g0}	E_{g0}	O_{g1}	E_{g1}	…	$O_{g,c-1}$	$E_{g,c-1}$

χοι καλής προσαρμογής έχουν αναπτυχθεί για το πολυωνυμικό μοντέλο λογιστικής παλινδρόμησης. Οι Fagerland (2009) [5] και Fagerland et al. (2008) [4] πρότειναν έναν έλεγχο καλής προσαρμογής για την περίπτωση που η μεταβλητή απόκρισης είναι κατηγορική με πάνω από δυο επίπεδα, ο οποίος αποτελεί γενίκευση του αντίστοιχου ελέγχου των Hosmer και Lemeshow $\hat{C}_g$ για το δυαδικό μοντέλο που αναφέρθηκε στο Κεφάλαιο 2. Η ομαδοποίηση των παρατηρήσεων βασίζεται και πάλι στην μέθοδο *«deciles of risk»* (βλ. Ενότητα 2.2), σύμφωνα με τις εκτιμώμενες πιθανότητες $1 - \pi_{i0} = \sum_{j=1}^{c-1} \pi_{ij}$.

Μετά την ομαδοποίηση των παρατηρήσεων και από την Σχέση (2.21) έπεται πως η ελεγχοσυνάρτηση $\hat{C}_g$ υπολογίζεται εφαρμόζοντας την στατιστική συνάρτηση του Pearson στον $g \times c$ Πίνακα 3.1 των παρατηρούμενων και αναμενόμενων συχνοτήτων.

$$\hat{C}_g = \sum_{k=1}^{g} \sum_{j=0}^{c-1} \frac{(O_{kj} - E_{kj})^2}{E_{kj}}. \tag{3.5}$$

Υπό την μηδενική υπόθεση ότι το μοντέλο προσαρμόζεται καλά στα δεδομένα, η ελεγχοσυνάρτηση $\hat{C}_g$ ακολουθεί ασυμπτωτικά της χ^2 κατανομή με $(g-2) \times (c-1)$ βαθμούς ελευθερίας [10].

Στην συνέχεια της παρούσας διπλωματικής θα αναπτυχθεί μια τροποποιημένη στρατηγική του παραπάνω ελέγχου, η οποία διαφοροποιείται ως προς τον τρόπο ομαδοποίησης. Το βασικό πλεονέκτημα της νέας στρατηγικής έγκειται στο γεγονός πως όταν το μοντέλο παρουσιάσει κακή προσαρμογή στα δεδομένα, τα άτομα (subjects) που δεν μοντελοποιούνται σωστά μπορούν να εντοπισθούν μέσω του Πίνακα Συνάφειας 3.1.

3.2 Έλεγχοι καλής προσαρμογής βασισμένοι σε γνωστές μεθόδους συσταδοποίησης

Οι έλεγχοι που θα αναπτυχθούν βασίζονται στην στρατηγική που προτάθηκε αρχικά από τον Tsiatis (1980) [16] για το διαχωρισμό του συνόλου των συμμεταβλητών $(x_1, x_2, \ldots, x_p)'$ σε g διακεκριμένες περιοχές στον p-διάστατο χώρο $R_1, R_2, \ldots, R_g$. Θα εστιάσουμε στο μοντέλο που περιέχει συνεχείς συμμεταβλητές, καθώς οι έλεγχοι Pearson και Απόκλισης ανταποκρίνονται καλά όταν στο μοντέλο περιέχονται μόνο κατηγορικές μεταβλητές. Στόχος είναι ο διαχωρισμός του συνόλου των συμμεταβλητών σε g διακεκριμένες περιοχές (ομάδες), σύμφωνα με γνωστές *μεθόδους Συσταδοποίησης (Clustering methods)* για την διεξαγωγή κλασικών X^2 ελέγχων καλής προσαρμογής. Η ιδέα της προσέγγισης αυτής αποτελεί άμεση γενίκευση της αντίστοιχης προσέγγισης των Xie et al. (2008) [19] για το δυαδικό λογιστικό μοντέλο, η οποία αργότερα επεκτάθηκε από τους Xie και Bian (2009) [20] για το διατακτικό λογιστικό μοντέλο. Η συγκεκριμένη στρατηγική προσφέρει σημαντικά πλεονεκτήματα. Αρχικά, αν ο έλεγχος υποδεικνύει έλλειψη καλής προσαρμογής του μοντέλου, είναι αρκετά εύκολο να ορισθούν οι τύποι των ατόμων που δεν μοντελοποιούνται σωστά [20]). Ένα πλεονέκτημα αποτελεί ακόμη, πως η συγκεκριμένη προσέγγιση μπορεί να χρησιμοποιηθεί για την αξιολόγηση της καλής προσαρμογής για όλους τους τύπους μοντέλων λογιστικής παλινδρόμησης - δυαδικό, πολυωνυμικό και διατακτικό μοντέλο [9].

3.2.1 Ανάλυση κατά συστάδες

Η ανάλυση κατά συστάδες (cluster analysis) [2], [26] αποσκοπεί στον διαχωρισμό μιας συλλογής από στοιχεία σε υποσύνολα, έτσι ώστε να υπάρχει ομοιογένεια μέσα σε ένα υποσύνολο και ανομοιογένεια μεταξύ των στοιχείων που ανήκουν σε διαφορετικά υποσύνολα. Ακόμη μπορεί να αποσκοπεί και στην ιεραρχική οργάνωση των συστάδων με την διαδοχική ομαδοποίηση αυτών, έτσι ώστε σε κάθε στάδιο της ιεραρχίας, οι συστάδες που ανήκουν στην ίδια ομάδα να είναι πιο όμοιες μεταξύ τους από αυτές που ανήκουν σε άλλη ομάδα. Για να μπορέσει να πραγματοποιηθεί συσταδοποίηση χρειάζονται κατάλληλα μέτρα, τα οποία θα μπορούν να υποδείξουν, πότε δύο παρατηρήσεις είναι όμοιες ή ανόμοιες μεταξύ τους. Τέτοια μέτρα είναι οι αποστάσεις (distances). Ως απόσταση ορίζεται το μέτρο, βάσει του οποίου δημιουργούνται οι συστάδες. Οι πιο διαδομένες μετρικές απόστασης στην βιβλιο-

γραφία είναι οι ακόλουθες : η μετρική Mahalanobis, Manhattan ή City – block, Minkowski της οποίας αποτελεί ειδική περίπτωση η Ευκλείδεια απόσταση, max ή Chebyshev, Gower και Canberra. Στην εργασία αυτή θα χρησιμοποιηθεί η Ευκλείδεια απόσταση. Επίσης, υπάρχουν αρκετές μέθοδοι συσταδοποίησης όπως είναι οι ιεραρχικές (hierarhical), οι διαχωριστικές (partitioning), μέθοδοι βασισμένες στην πυκνότητα (density-based), βασισμένες στο μοντέλο (model-based) και άλλες. Θα αναφερθούμε σε ορισμένες από τις διαχωριστικές και ιεραρχικές μεθόδους συσταδοποίησης.

- **Ιεραρχική συσσωρευτική συσταδοποίηση (Hierarhical agglomerative clustering)**, κατά την οποία οι συστάδες σχηματίζονται σταδιακά, παράγοντας μια ιεραρχία δενδροειδούς μορφής. Σε αυτή την μέθοδο συσταδοποίησης δεν υπολογίζεται η απόσταση μεταξύ των στοιχείων, αλλά η απόσταση μεταξύ των συστάδων που έχουν προέλθει είτε από την συγχώνευση άλλων συστάδων είτε από την συγχώνευση άλλων παρατηρήσεων. Υπάρχουν αρκετές αντιπροσωπευτικές μέθοδοι για τον προσδιορισμό των αποστάσεων στην συγκεκριμένη κατηγορία, μεταξύ αυτών, η μέθοδος της απλής συνένωσης (single linkage method), η μέθοδος της πλήρους συνένωσης (complete linkage method), η μέθοδος των σταθμισμένων μέσων (weighted average linkage method), η μέθοδος των κέντρων βάρους (centroid method) και η μέθοδος του Ward (Ward's method). Στην εργασία αυτή θα χρησιμοποιηθεί η μέθοδος του Ward, η οποία δεν υπολογίζει αποστάσεις μεταξύ των συστάδων αλλά προσδοκεί την ελαχιστοποίηση της διακύμανσης μέσα σε αυτές, ενώ έχει την τάση να δημιουργεί συστάδες με παρόμοιο αριθμό παρατηρήσεων. Γενικότερα οι ιεραρχικές μέθοδοι δεν ενδείκνυνται για μεγάλο πλήθος δεδομένων, καθώς απαιτούν πολύ χρόνο, μνήμη και ισχύ.

- **K-means συσταδοποίηση** χρησιμοποιείται για τον διαχωρισμό ενός συνόλου δεδομένων σε k-συστάδες, όπου το πλήθος των συστάδων k καθορίζεται κάθε φορά από τον ερευνητή. Στόχος της μεθόδου είναι η κατανομή ενός συνόλου αντικειμένων σε ένα προκαθορισμένο αριθμό συστάδων k με τρόπο τέτοιο ώστε να αυξάνει την ομοιότητα των στοιχείων εντός των συστάδων και ταυτόχρονα να ελαχιστοποιεί την συνολική διακύμανση εντός αυτών. Ο αλγόριθμος περιλαμβάνει μια επαναληπτική διαδικασία, όπου σε κάθε επανάληψη υπολογίζεται το κέντρο της συστάδας (centroid). Τα αντικείμενα εντάσσονται στη συστάδα με το πλησιέστερο κέντρο. Στην μέθοδο k-means το κέντρο της ομάδας είναι στην πραγματικότητα η μέση τιμή των σημείων που ανα-

τίθενται στην συστάδα. Είναι απλή και γρήγορη μέθοδος και μπορεί να διαχειριστεί μεγάλα σύνολα δεδομένων. Παρόλ' αυτά επηρεάζεται σημαντικά από τις αρχικές τιμές των κεντρών των συστάδων καθώς επίσης και από την ύπαρξη θορύβου (noise) και έκτροπων παρατηρήσεων (outliers) στα δεδομένα. Αν κατά την εκτέλεσή του βρεθεί αντιμέτωπος με κάποια έκτροπη παρατήρηση, τότε αναπόφευκτα θα την αναθέσει σε κάποια συστάδα, επηρεάζοντας άμεσα την τιμή του αντίστοιχου κέντρου κι επομένως την ποιότητα της τελικής συσταδοποίησης.

- **PAM (Partitioning Around Medoids) συσταδοποίηση** ανήκει στην υποκατηγορία των k-medoids μεθόδων. Στόχος της μεθόδου, όπως και στην k-means, είναι ο διαχωρισμός ενός συνόλου σε k-συστάδες, όπου κάθε στοιχείο θα βρίσκεται στη συστάδα με το κοντινότερο κεντροειδές (medoid). Η διαφορά των δυο μεθόδων έγκειται στην επιλογή του κέντρου κάθε συστάδας. Ως *«medoid»* ορίζεται ένα στοιχείο μέσα στην συστάδα για το οποίο η μέση ανομοιότητα αυτού με τα στοιχεία της συστάδας ελαχιστοποιείται. Με άλλα λόγια, το medoid είναι το στοιχείο που βρίσκεται πιο κεντρικά τοποθετημένο στην συστάδα. Η PAM χρησιμοποιεί δυο μετρικές απόστασης : την Ευκλείδεια και την απόσταση Manhattan. Στην πράξη, η χρήση της μιας ή της άλλης μετρικής οδηγεί συνήθως σε παρόμοια αποτελέσματα. Όμως στην περίπτωση που στα δεδομένα υπάρχουν έκτροπες παρατηρήσεις, η απόσταση Manhattan δίνει πιο ισχυρά αποτελέσματα, συγκριτικά με την Ευκλείδεια η οποία θα επηρεαζόταν από αυτές τις τιμές. Επιπλέον η PAM αποτελεί πιο ισχυρή μέθοδος από την k-means, καθώς έχει την ικανότητα να χειρίζεται καλά τις έκτροπες παραηρήσεις. Ακόμη, η διάταξη της εισόδου δεν επηρεάζει τα αποτελέσματα, όπως συμβαίνει με την k-means. Τέλος, η PAM δουλεύει ικανοποιητικά για μικρά σύνολα δεδομένων, ενώ αποδεικνύεται μη αποδοτική για σύνολα δεδομένων μεσαίου και μεγάλου μεγέθους, λόγω της μεγάλης πολυπλοκότητάς της.

Πίνακας 3.2: *Περιγραφή των ελεγχοσυναρτήσεων*

	Ελεγχοσυνάρτηση	**Κατηγορία Συσταδοποίησης (μέθοδος)**
1	X^2_{w*g}	Ιεραρχική (μέθοδος Ward)
2	X^2_{k*g}	Διαχωριστική (μέθοδος k-means)
3	X^2_{p*g}	Διαχωριστική (μέθοδος PAM)

Κεφάλαιο 4

Προσομοίωση

4.1 Σχεδιασμός προσομοίωσης

Οι προσομοιώσεις πραγματοποιούνται με χρήση της γλώσσας προγραμματισμού ανοικτού κώδικα R. Οι παράμετροι που επιτρέπονται κάθε φορά να μεταβάλλονται είναι : το μέγεθος δείγματος n, ο τύπος του αποκλινόμενου μοντέλου από το πραγματικό μοντέλο, οι συντελεστές παλινδρόμησης β και η κατανομή της επεξηγηματικής μεταβλητής που παραλείπεται. Ο σχεδιασμός της προσομοίωσης ακολουθεί πιστά τη γραμμή των Fagerland et al. (2008) [4]. Το μοντέλο που χρησιμοποιείται, για λόγους ευκολίας, περιέχει μόνο μια συνεχή επεξηγηματική μεταβλητή, ενώ η αποκρινόμενη Y θεωρείται ότι έχει τρία επίπεδα ($c = 3$). Προφανώς προκύπτουν δυο logit μετασχηματισμοί, έστω g_1 και g_2. Επιπλέον για εξοικονόμηση χώρου και υπολογιστικού χρόνου, από την μελέτη των Fagerland et al. θα αναφερθούμε μόνο : i) στο μοντέλο με συντελεστές παλινδρόμησης $\beta_{10} = -2.10$, $\beta_{11} = -0.35$, $\beta_{20} = -1.90$, $\beta_{21} = -0.21$, ii) στις κατανομές $N(0, 3)$ και $\chi^2(4)$ για την επεξηγηματική x και τέλος iii) θα εξετάσουμε μόνο την περίπτωση όπου $g = 10$ ομάδες.

4.1.1 Βήματα Προσομοίωσης

Για κάθε μέγεθος δείγματος κατασκευάζεται το μοντέλο με μια επεξηγηματική μεταβλητή x για τις τιμές του $\boldsymbol{\beta}$ που δίνονται παραπάνω. Ακολουθείται η εξής διαδικασία [4] :

i) παραγωγή τυχαίων τιμών από συγκεκριμένη κατανομή για την επεξηγηματική x

ii) προσαρμογή των πολυωνυμικών logit μετασχηματισμών $g_1(x) = \beta_{10}+\beta_{11}x$ και $g_2(x) =$

$\beta_{11} + \beta_{21}x$ στα προσομοιωμένα δεδομένα

iii) υπολογισμός των $g_1(x)$ και $g_2(x)$ και κατόπιν υπολογισμός των πιθανοτήτων π_0, π_1, π_2 από τις Σχέσεις (3.2) και (3.3).

iv) έστω $u \sim U(0, 1)$

v) παραγωγή των τιμών της αποκρινόμενης Y με χρήση των εκτιμώμενων πιθανοτήτων του πολυωνυμικού λογιστικού μοντέλου, σύμφωνα με τον κανόνα που χρησιμοποίησαν και οι Fagerland et al. [4] που υποδεικνύει ότι i) $y = 2$ αν $u > \pi_0 + \pi_1$, ii) $y = 1$ αν $u < \pi_0 + \pi_1$ και $u > \pi_0$, iii) $y = 0$ διαφορετικά

vi) προσαρμογή του πολυωνυμικού μοντέλου λογιστικής παλινδρόμησης

vii) υπολογισμός των $\hat{C}_g, X^2_{w*g}, X^2_{k*g}, X^2_{p*g}$ για $g = 10$

viii) υπολογισμός των ρυθμών απόρριψης (RR) του μοντέλου για κάθε έλεγχο.

Πραγματοποιήθηκαν 10.000 προσομοιώσεις για κάθε κατανομή ($N(0, 3)$,$\chi^2(4)$) και για κάθε μέγεθος δείγματος ($n = 100, 400$). Ο κώδικας της προσομοίωσης [27] δίνεται από τους Hamid et al. (2017) [9] και Hamid et al. (2018) [8].

4.2 Προσομοίωση για το πραγματικό μοντέλο

Οι προσομοιωμένοι ρυθμοί απόρριψης σε $\alpha = 5\%$ επίπεδο σημαντικότητας, όταν προσαρμόζεται το πραγματικό πολυωνυμικό λογιστικό μοντέλο με μια συνεχή επεξηγηματική μεταβλητή δίνονται στον Πίνακα 4.1. Τα αποτελέσματα δείχνουν πως οι ρυθμοί απόρριψης και για τους τέσσερις ελέγχους είναι πολύ κοντά στο 5% ε.σ., ενώ για $n = 400$ τα αποτελέσματα είναι καλύτερα σε σχέση με το μικρότερο μέγεθος δείγματος. Επιπλέον, παρατηρείται πως για την κατανομή $\chi^2(4)$, οι ρυθμοί απόρριψης των ελέγχων για μέγεθος δείγματος 100 επηρεάζονται αισθητά από την υψηλή ασυμμετρία της. Για $n = 400$ η επιρροή είναι μικρότερη και προκύπτουν καλύτερα αποτελέσματα, δηλαδή τιμές πιο κοντά στο ονομαστικό επίπεδο σημαντικότητας $\alpha = 5\%$.

Πίνακας 4.1: *Ρυθμοί απόρριψης σε 5% ε.σ. για το πραγματικό μοντέλο.*

	$n = 100$				$n = 400$			
	$\hat{C}_{10}$	X^2_{w*10}	X^2_{k*10}	X^2_{p*10}	$\hat{C}_{10}$	X^2_{w*10}	X^2_{k*10}	X^2_{p*10}
$\boldsymbol{N(0,3)}$	4.62	5.56	4.65	5.5	4.91	6.03	5.07	5.13
$\boldsymbol{\chi^2(4)}$	1.95	1.50	2.21	0.97	4.15	4.32	4.19	4.24

4.3 Ισχύς των ελέγχων

Στην συνέχεια θα εξετασθεί η ισχύς των παραπάνω ελέγχων. Η ισχύς θα υπολογισθεί σύμφωνα με την ικανότητα κάθε ελέγχου να εντοπίσει τυχούσες αποκλίσεις από το πραγματικό μοντέλο. Για τον λόγο αυτό, θεωρούνται τρεις πιθανές καταστάσεις :

i) η παράλειψη ενός τετραγωνικού όρου (omission of a quadratic term)

ii) η παράλειψη ενός όρου αλληλεπίδρασης μεταξύ δυο συνεχών επεξηγηματικών μεταβλητών (omission of a continuous interaction term)

iii) η παράλειψη ενός όρου αλληλεπίδρασης μεταξύ μια συνεχούς και μιας δυαδικής επεξηγηματικής μεταβλητής (omission of a dichotomous interaction term).

Η συγκεκριμένη μελέτη έχει γίνει και από τους Hamid et al. (2017) [9] και Hamid et al. (2018) [8]. Οι Hamid et.al εξέτασαν την ισχύ των ελέγχων $\hat{\boldsymbol{C}}_g$ και X^2_{w*g}, ενώ διεξήγαγαν συγκρίσεις και συμπεράσματα μεταξύ αυτών για καθεμία από τις παραπάνω καταστάσεις. Επίσης, διεξήγαγαν συγκρίσεις και συμπεράσματα μεταξύ των X^2_{w*g} και X^2_{k*g} για κάθε κατάσταση. Τα αποτελέσματα είναι άμεσα συγκρίσιμα μεταξύ τους, καθώς χρησιμοποιείται ο ίδιος σχεδιασμός προσομοίωσης. Στην παρούσα διπλωματική, προτείνουμε την μέθοδο συσταδοποιήσης PAM για τον διαχωρισμό του συνόλου των συμμεταβλητών σε g ομάδες για τον ίδιο σχεδιασμό προσομοίωσης. Στόχος είναι να ερευνηθεί εαν άλλες μέθοδοι συσταδοποίησης, όπως είναι η PAM, θα βελτιώσει την ισχύ του προτεινόμενου ελέγχου καλής προσαρμογής. Παρακάτω απεικονίζονται σε κοινούς πίνακες οι ρυθμοί απόρριψης των τεσσάρων ελέγχων για τις τρεις παραπάνω υποθετικές καταστάσεις απόκλισης του προσαρμοσμένου μοντέλου από το πραγματικό μοντέλο. Υπενθυμίζεται πως μεγάλος ρυθμός απόρριψης συνεπάγεται μεγάλη ισχύς του ελέγχου.

4.3.1 Παράλειψη ενός τετραγωνικού όρου

Για να εξετάσουμε την ισχύ των τεσσάρων ελέγχων καλής προσαρμογής, εξετάζουμε την ισχύ καθενός για τον εντοπισμό ενός τετραγωνικού όρου που παραλείπεται από το πραγματικό μοντέλο. Οι τιμές της αποκρινόμενης Y παράγονται από ένα μοντέλο με logits

$$g_j(x) = \beta_{j0} + \beta_{j1}x + \beta_{j2}x^2, \quad \text{για} \quad j = 1, 2.$$

Πίνακας 4.2: *Ρυθμοί απόρριψης σε 5% ε.σ. για την παράλειψη τετραγωνικού όρου.*

	$n = 100$				$n = 400$			
β_{j2}	$\hat{C}_{10}$	X^2_{w*10}	X^2_{k*10}	X^2_{p*10}	$\hat{C}_{10}$	X^2_{w*10}	X^2_{k*10}	X^2_{p*10}
$\boldsymbol{N(0,3)}$								
0.01	5.55	8.73	7.78	8.13	7.28	11.97	10.85	9.28
0.05	19.00	40.26	33.75	35.09	70.79	87.92	86.15	82.69
0.10	55.65	80.47	75.93	77.03	99.98	100	99.99	100
0.20	89.17	98.10	97.69	97.89	100	100	100	100
0.30	94.11	99.64	99.57	99.64	100	100	100	100
0.40	95.54	99.82	99.90	99.87	100	100	100	100
0.50	96.09	99.88	99.91	99.91	100	100	100	100
$\boldsymbol{\chi^2(4)}$								
0.01	2.73	3.46	3.42	2.13	5.13	11.21	8.31	8.15
0.05	11.30	9.99	12.71	10.08	38.69	37.09	41.32	40.51
0.10	12.88	8.45	13.18	9.29	45.77	31.82	40.89	37.97
0.20	7.24	4.96	8.82	5.10	30.00	16.40	22.66	21.41
0.30	4.29	3.34	6.36	3.36	18.06	8.67	14.89	10.67
0.40	2.78	2.48	4.39	2.21	10.58	4.81	9.15	6.30
0.50	1.78	1.60	3.26	1.53	6.56	3.11	60.20	3.58

* $\beta_{10} = -2.10$, $\beta_{11} = -0.35$, $\beta_{20} = -1.90$, $\beta_{21} = -0.21$.

Το προσαρμοσμένο μοντέλο περιέχει μόνο τους δυο πρώτους όρους, δηλαδή $b_{j2} = 0$. Οι συντελεστές παλινδρόμησης β_{j2} παίρνουν τις τιμές 0.01, 0.05, 0.1, 0.2, 0.3, 0.4 και 0.5

που αντιστοιχούν στην αυξανόμενη διαφορά μεταξύ των δυο μοντέλων (με και χωρίς τον τετραγωνικό όρο). Από τον Πίνακα 4.2 φαίνεται πως όλοι οι έλεγχοι έχουν ικανοποιητική ισχύ για την κατανομή $N(0,3)$. Ειδικά για μεγαλύτερο μέγεθος δείγματος ($n = 400$) προκύπτει η επιθυμητή ισχύς. Επιπλέον, όσο αυξάνει η διαφορά μεταξύ των μοντέλων, δηλαδή οι τιμές της β_{j2}, τόσο αυξάνει και η ισχύς των ελέγχων. Παρόλ'αυτά κάτι τέτοιο δεν ισχύει για την κατανομή $\chi^2(4)$. Λόγω της υψηλής ασυμμετρίας της, οι ρυθμοί απόρριψης είναι πολύ χαμηλοί και για τα δυο μεγέθη δείγματος, με αποτέλεσμα και οι τέσσερις έλεγχοι να καθίστανται μη αξιόπιστοι. Σε γενικές γραμμές, οι έλεγχοι φαίνεται να έχουν παρόμοια ισχύ για την συγκεκριμένη περίπτωση.

4.3.2 Παράλειψη ενός συνεχούς όρου αλληλεπίδρασης

Κατόπιν, εξετάζουμε την ισχύ κάθε ελέγχου να εντοπίσει έναν όρο αλληλεπίδρασης που παραλείπεται από το πραγματικό μοντέλο μεταξύ δυο συνεχών επεξηγηματικών μεταβλητών x_1 και x_2. Οι τιμές της Y παράγονται από ένα μοντέλο με logit μετασχηματισμούς

$$g_j(x) = \beta_{j0} + \beta_{j1}x_1 + \beta_{j2}x_2 + \beta_{j3}x_1x_2, \quad \text{για} \quad j = 1, 2.$$

Αρχικά, η προσαρμογή του μοντέλου έγινε μόνο με τους τρεις πρώτους όρους, χωρίς τον όρο αλληλεπίδρασης ($b_{j3} = 0$). Όσον αφορά τις κατανομές των συμμεταβλητών του μοντέλου, η κατανομή της x_1 επιτρέπεται να μεταβάλλεται μεταξύ των $N(0,3)$ και $\chi^2(4)$, ενώ η κατανομή της x_2 είναι σταθερά η $N(0,1)$. Για να διερευνηθεί η επίδραση των αυξανόμενων επιπέδων αλληλεπίδρασης για διαφορετικές τιμές των συντελεστών του συνεχούς όρου, χρησιμοποιήθηκε ένας συνδυασμός τριων τιμών της β_{j2} (0.2, 0.6, 1.0) και τριων τιμών της β_{j3} (0.2, 0.6, 1.0). Οι ρυθμοί απόρριψης των ελέγχων σε 5% ε.σ. δίνονται στον Πίνακα 4.3. Τα αποτελέσματα των προσομοιώσεων έδειξαν πως ο X^2_{w*10} έχει αισθητά μεγαλύτερη ισχύ σε σχέση με τους $\hat{\boldsymbol{C}}_{10}$ και X^2_{k*10}, αλλά όχι ιδιαίτερα μεγάλη διαφορά με τον X^2_{p*10}. Παρατηρείται πως τα ποσοστά των X^2_{w*10} και X^2_{p*10} δεν διαφέρουν σημαντικά, δηλαδή έχουν παρόμοια ισχύ, όταν η κατανομή της επεξηγηματικής x_1 είναι η $N(0,3)$ και για τα δυο μεγέθη δείγματος. Όταν η κατανομή της x_1 είναι η $\chi^2(4)$ παρατηρείται πως όλοι οι έλεγχοι επηρεάζονται από την υψηλή ασυμμετρία της κατανομής και έχουν χαμηλότερη ισχύ, με τον $\hat{\boldsymbol{C}}_{10}$ να εμφανίζει την καλύτερη ισχύ μεταξύ αυτών. Ωστόσο και για αυτή την περίπτωση,

ο X^2_{p*10} φαίνεται να μην διαφέρει σηματικά από τον $\hat{\boldsymbol{C}}_{10}$ και να είναι ο αμέσως επόμενος με την μεγαλύτερη ισχύ κατά πλειοψηφία. Επιπλέον, όσο αυξάνει το δείγμα, αυξάνει και η ισχύς των $\hat{\boldsymbol{C}}_{10}$ και X^2_{p*10}. Ακόμη, παρατηρείται πως η ισχύς όλων των ελέγχων αυξάνει με την αύξηση των επιπέδων αλληλεπίδρασης.

Πίνακας 4.3: *Ρυθμοί απόρριψης σε 5% ε.σ. για την παράλειψη συνεχούς όρου αλληλεπίδρασης*

		$n = 100$				$n = 400$			
β_{j2}	β_{j3}	$\hat{C}_{10}$	X^2_{w*10}	X^2_{k*10}	X^2_{p*10}	$\hat{C}_{10}$	X^2_{w*10}	X^2_{k*10}	X^2_{p*10}
$\boldsymbol{N(0,3)}$									
	0.2	5.56	9.48	7.25	7.89	6.82	23.19	9.53	18.48
0.2	0.6	18.96	42.74	10.79	38.59	49.89	95.71	25.28	93.70
	1.0	35.41	68.30	12.28	65.34	76.10	99.77	35.48	99.71
	0.2	4.50	8.16	6.53	6.66	11.15	21.15	9.39	16.08
0.6	0.6	12.17	40.82	10.10	37.14	57.79	95.42	22.46	92.81
	1.0	27.05	66.99	11.63	64.57	75.92	99.72	31.32	99.66
	0.2	3.73	6.76	5.38	5.21	9.83	15.96	7.78	11.81
1.0	0.6	10.43	37.50	8.87	33.20	67.09	93.14	18.67	89.15
	1.0	21.25	65.01	10.97	61.75	78.28	99.52	26.78	99.49
$\boldsymbol{\chi^2(4)}$									
	0.2	3.07	3.17	0.13	2.51	4.44	6.75	5.17	7.22
0.2	0.6	12.31	8.29	2.35	9.17	23.90	22.43	1.28	29.11
	1.0	16.15	9.49	1.98	11.27	39.49	26.49	0.84	35.41
	0.2	3.75	3.25	1.03	3.53	4.95	5.25	3.42	5.68
0.6	0.6	11.11	6.90	1.81	24.44	22.44	15.57	0.98	20.93
	1.0	14.28	8.08	1.47	8.71	32.77	19.15	0.68	26.24
	0.2	5.34	3.97	2.21	3.90	6.23	4.58	2.31	5.42
1.0	0.6	10.20	6.38	1.54	7.13	20.49	11.62	1.01	16.00
	1.0	12.71	7.53	1.25	7.53	26.95	14.15	0.40	20.09

* $\beta_{10} = -2.10$, $\beta_{11} = -0.35$, $\beta_{20} = -1.90$, $\beta_{21} = -0.21$.

4.3.3 Παράλειψη ενός συνεχούς*δυαδικής όρου αλληλεπίδρασης

Τέλος, θα εξετάσουμε την ισχύ των ελέγχων για τον εντοπισμό ενός όρου αλληλεπίδρασης που παραλείπεται μεταξύ μιας συνεχούς επεξηγηματικής x και μια κατηγορικής μεταβλητής d με δυο επίπεδα, όπου η d ακολουθεί την κατανομή $Bernoulli(1, \frac{1}{2})$. Οι τιμές της αποκρινόμενης Y παράγονται από ένα μοντέλο με logit μετασχηματισμούς

$$g_j(x) = \beta_{j0} + \beta_{j1}x_+\beta_{j2}d + \beta_{j3}xd, \quad \text{για} \quad j = 1, 2.$$

Πίνακας 4.4: *Ρυθμοί απόρριψης σε 5% ε.σ. για την παράλειψη δυαδικού όρου αλληλεπίδρασης*

		$n = 100$				$n = 400$			
β_{j2}	β_{j3}	$\hat{C}_{10}$	X^2_{w*10}	X^2_{k*10}	X^2_{p*10}	$\hat{C}_{10}$	X^2_{w*10}	X^2_{k*10}	X^2_{p*10}
$\boldsymbol{N(0,3)}$									
	0.2	4.60	4.89	5.47	3.93	6.17	6.17	6.17	5.78
0.2	0.6	20.31	13.00	7.85	14.06	79.12	30.39	13.83	54.76
	1.0	51.20	28.37	13.51	36.42	97.48	67.78	32.71	95.69
midrule	0.2	5.03	4.53	5.68	3.90	8.36	5.84	5.65	5.77
0.6	0.6	25.68	12.96	7.22	14.07	91.87	28.22	10.64	54.72
	1.0	61.22	27.61	11.62	36.51	99.74	64.33	25.90	95.50
	0.2	4.52	4.36	4.89	3.33	12.74	5.75	5.47	5.50
1.0	0.6	31.95	11.91	5.95	13.51	94.51	25.34	8.00	53.64
	1.0	70.00	26.23	10.23	35.29	100	60.96	19.81	94.96
$\boldsymbol{\chi^2(4)}$									
	0.2	2.46	1.96	3.41	1.66	5.72	4.83	4.98	3.93
0.2	0.6	19.03	7.72	2.29	9.39	84.45	29.95	1.61	50.84
	1.0	38.45	16.38	1.44	21.09	99.53	61.71	3.60	90.31
	0.2	2.43	1.92	3.09	1.93	5.48	4.60	4.60	3.84
0.6	0.6	17.09	6.35	1.81	7.91	81.98	26.22	1.24	47.10
	1.0	29.27	11.56	1.02	15.55	98.31	52.32	2.19	83.40
	0.2	2.31	1.88	2.67	1.74	5.45	4.42	3.97	3.49
1.0	0.6	14.38	5.07	1.34	6.54	76.77	21.44	0.87	40.47
	1.0	19.97	7.91	0.80	10.55	94.10	41.77	1.20	73.25

* $\beta_{10} = -2.10$, $\beta_{11} = -0.35$, $\beta_{20} = -1.90$, $\beta_{21} = -0.21$.

Ακολουθείται και πάλι η ίδια διαδικασία με παραπάνω, δηλαδή προσαρμόζεται το πολυωνυμικό λογιστικό μοντέλο χωρίς τον όρο αλληλεπίδρασης ($b_{j3} = 0$), ενώ λαμβάνεται πάλι ο ίδιος συνδυασμός τιμών των συντελεστών παλινδρόμησης β_{j2} και β_{j3} που αναφέραμε παραπάνω για την παράλειψη ενός συνεχούς όρου αλληλεπίδρασης. Τα αποτελέσματα των ρυθμών απόρριψης σε 5% ε.σ. δίνονται στον Πίνακα 4.4. Είναι φανερό πως ο $\hat{C}_{10}$ έχει την μεγαλύτερη ισχύ μεταξύ των υπολοίπων τόσο για τις δυο κατανομές της επεξηγηματικής x, όσο και για τα δυο μεγέθη δείγματος. Παρόλ' αυτά, ο X^2_{p*10} για μεγαλύτερο μέγεθος δείγματος φαίνεται να έχει και αυτός μεγάλη ισχύ, άμεσα συγκρίσιμη με του $\hat{C}_{10}$. Παρατηρείται ακόμη μείωση της ισχύς των ελέγχων για τα μοντέλα με επεξηγηματική μεταβλητή την $\chi^2(4)$. Όπως και στις παραπάνω περιπτώσεις, όλοι οι έλεγχοι έχουν μεγαλύτερη ισχύ για μεγαλύτερα επίπεδα αλληλεπίδρασης.

4.4 Συμπεράσματα Προσομοίωσης

Στο κεφάλαιο αυτό μελετήσαμε, μέσω προσομοίωσης, την ισχύ ελέγχων καλής προσαρμογής που βασίζονται στο διαχωρισμό του συνόλου των επεξηγηματικών μεταβλητών σε ομάδες, σύμφωνα με γνωστές μεθόδους συσταδοποίησης και παράλληλα διεξήγαμε συγκρίσεις μεταξύ αυτών με τον γνωστό HL έλεγχο καλής προσαρμογής, ο οποίος βασίζεται στην ομαδοποίηση των εκτιμώμενων πιθανοτήτων του προσαρμοσμένου μοντέλου, για την περίπτωση του πολυωνυμικού λογιστικού μοντέλου. Υπενθυμίζεται πως ο έλεγχος HL πέραν του ότι επηρεάζεται τόσο από το μέγεθος δείγματος όσο και από το πλήθος των ομάδων που δημιουργούνται, αδυνατεί να εντοπίσει τα άτομα που δεν μοντελοποιούνται σωστά, όταν το μοντέλο υποδεικνύει κακή προσαρμογή στα δεδομένα, σε αντίθεση με τους προτεινόμενους ελέγχους που έχουν τη δυνατότητα αυτή μέσω του πίνακα συνάφειας. Οι Hamid et.al (2017) [9] χρησιμοποίησαν την μέθοδο Ward από τις ιεραρχικές μεθόδους και οι Hamid et.al (2018) [8] την k-means από τις διαχωριστικές μεθόδους για τον διαχωρισμό του συνόλου των συμμεταβλητών σε g ομάδες και τον ίδιο σχεδιασμό προσομοίωσης. Στις μελέτες τους παροτρύνουν την χρήση διαφορετικών μεθόδων συσταδοποίησης για την βελτίωση της ισχύς των ελέγχων αυτών. Στην πραγματικότητα, η μελέτη των ελέγχων καλής προσαρμογής του λογιστικού μοντέλου δεν τελειώνουν εδώ, καθώς μπορούν να χρησιμοποιηθούν και άλλες τεχνικές συσταδοποιήσης για τον διαχωρισμό των ομάδων και στην συνέχεια την εφαρμογή ενός κλασικού X^2 ελέγχου καλής προσαρμογής του Pearson. Έτσι βασιζόμενοι στα απο-

τελέσματα αυτών, συμπεριλάβαμε στην ανάλυση και την μέθοδο PAM για την διερεύνηση της ισχύς της καθώς επίσης και για την σύγκριση αυτής με τα αποτελέσματα των προαναφερθέντων ερευνητών. Τα αποτελέσματα από τις προσομοιώσεις έδειξαν πως ο X^2_{p*10} έχει καλύτερη ισχύ σε σχέση με τον $\hat{C}_{10}$ να εντοπίσει τυχούσες αποκλίσεις από τον πραγματικό μοντέλο, όταν η επεξηγηματική μεταβλητή δεν παρουσιάζει υψηλή ασυμμετρία. Επιπλέον, η ισχύς του X^2_{p*10} (όπως και των υπολοίπων ελέγχων) αυξάνει, τόσο με την αύξηση του μεγέθους δείγματος, όσο και με την αύξηση της διαφοράς μεταξύ του μοντέλου που περιέχει τον όρο αλληλεπίδαρασης και του μοντέλου που δεν τον περιέχει.

Αξίζει να σημειωθεί ακόμη πως οι συγκεκριμένες μελέτες είναι περιορισμένες ως προς το πεδίο εφαρμογής τους, καθώς χρησιμοποιούν μόνο μια συνεχή επεξηγηματική μεταβλητή, τρεις κατηγορίες για την αποκρινόμενη μεταβλητή Y και δυο μεγέθη δείγματος. Ακόμη, στην εργασία αυτή χρησιμοποιήθηκαν μόνο $g = 10$ ομάδες για την κατασκευή των ελέγχων. Διαφορετική επιλογή του πλήθος των ομάδων να οδηγούσε πιθανώς σε διαφορετικές αποφάσεις για τον $\hat{\boldsymbol{C}}_g$, αφού σύμφωνα με την βιβλιογραφία, ένα από τα βασικά του μειονεκτήματα αποτελεί η δυσκολία επιλογής βέλτιστου αριθμού ομάδων.

Τέλος, σημαντική αποτέλεσε η συμβολή των Fagerland et al. (2012) [6], Hussain et al. (2015) [11], Pulkstenis et al. (2002) [13], Pulkstenis et al. (2004) [14] στη διαδικασία της προσομοίωσης.

Κεφάλαιο 5

Εφαρμογή

Στο κεφάλαιο αυτό θα εφαρμόσουμε τους ελέγχους καλής προσαρμογής που παρουσιάστηκαν στα Κεφάλαια 3 και 4 σε ένα πραγματικό σύνολο δεδομένων. Τα δεδομένα που θα χρησιμοποιηθούν είναι διαθέσιμα στην ιστοσελίδα https://www.kaggle.com/giripujar/hr-analytics. Το συγκεκριμένο σύνολο δεδομένων προέρχεται από την έρευνα του τμήματος Ανθρώπινου Δυναμικού για τους υπαλλήλους μιας μεγάλης εταιρείας. Αποτελείται από τις απαντήσεις 4410 υπαλλήλων, εκ των οποίων 9 υπάλληλοι παραλείπονται από την ανάλυση (omitted values) και 16 επεξηγηματικές μεταβλητές. Ένα ενδιαφέρον ερώτημα που θα μπορούσε να τεθεί για το συγκεριμένο σύνολο δεδομένων θα ήταν το εξής : ”Πόσο συχνά ένας υπάλληλος της εταιρείας ταξιδεύει για επαγγελματικούς σκοπούς ”, με βάση κάποια συγκεκριμένα χαρακτηριστικά, όπως η οικογενειακή του κατάσταση, το τμήμα της εταιρείας που εργάζεται και τα συνολικά χρόνια προϋπηρεσίας του. Στον Πίνακα 5.1 γίνεται περιγραφή των μεταβλητών που θα χρησιμοποιηθούν στην εφαρμογή αυτή. Η μεταβλητή απόκρισης BusinessTravel είναι κατηγορική με τρία επίπεδα (Ποτέ-Συχνά-Σπάνια). Συνεπώς, η χρήση της πολυωνυμικής λογιστικής παλινδρόμησης απαιτείται για την πρόβλεψη της συχνότητας των επαγγελματικών ταξιδιών κάθε υπαλλήλου. Από το δείγμα των 4410 υπαλλήλων, μελετήθηκαν 4401 άτομα, εκ των οποίων 449 (10.2%) υπάλληλοι δεν ταξιδεύουν ποτέ, 829 (18.8%) ταξιδεύουν συχνά ενώ 3123 (71%) υπάλληλοι ταξιδεύουν σπάνια για επαγγελματικούς σκοπούς. Αξίζει να σημειωθεί πως η αναλογία των παρατηρήσεων μεταξύ των κατηγοριών παίζει σημαντικό ρόλο, καθώς το μοντέλο θα προβλέπει συνέχεια την μεγαλύτερη κατηγορία, η οποία θα είναι πιο πιθανή να συμβεί. Όσο πιο άνισα κατανεμημένες είναι οι παρατηρήσεις σε κάθε κατηγορία της αποκρινόμενης μεταβλητής, τόσο πιο δύσκολη είναι η

πρόβλεψη της μικρότερης κατηγορίας αυτής. Επομένως, παρατηρείται κάποιου είδους «μεροληψία» των κατηγοριών της αποκρινόμενης BusinessTravel, η οποία θα πρέπει να ληφθεί υπόψιν στην ανάλυση. Παρακάτω δίνονται κάποια περιγραφικά στατιστικά και γραφήματα των επεξηγηματικών μεταβλητών σε σχέση με την μεταβλητή απόκρισης.

Πίνακας 5.1: *Περιγραφή των μεταβλητών του μοντέλου.*

Μεταβλητή	Τύπος	Τιμές	Περιγραφή
BusinessTravel (Αποκρινόμενη)	Κατηγορική	0=Ποτέ 1=Συχνά 2=Σπάνια	Πόσο συχνά ταξιδεύει ένας υπάλληλος της εταιρείας
MaritalStatus (Επεξηγηματική)	Κατηγορική	0=Χωρισμένος/η 1=Παντρεμένος/η 3=Ελεύθερος/η	Η οικογενειακή κατάσταση του υπαλλήλου
Department (Επεξηγηματική)	Κατηγορική	0= Τμήμα Ανθρώπινου Δυναμικού 1=Τμήμα Έρευνας και Ανάπτυξης 2=Τμήμα Πωλήσεων	Τμήμα της εταιρείας που εργάζεται ο υπάλληλος
TotalWorkingYears (Επεξηγηματική)	Συνεχής	0-40 χρόνια	Συνολικά χρόνια προϋπηρεσίας του υπαλλήλου

5.1 Περιγραφικά στατιστικά του μοντέλου

Πίνακας 5.2: *Συχνότητα ταξιδιών ως προς τα συνολικά έτη προϋπηρεσίας των υπαλλήλων.*

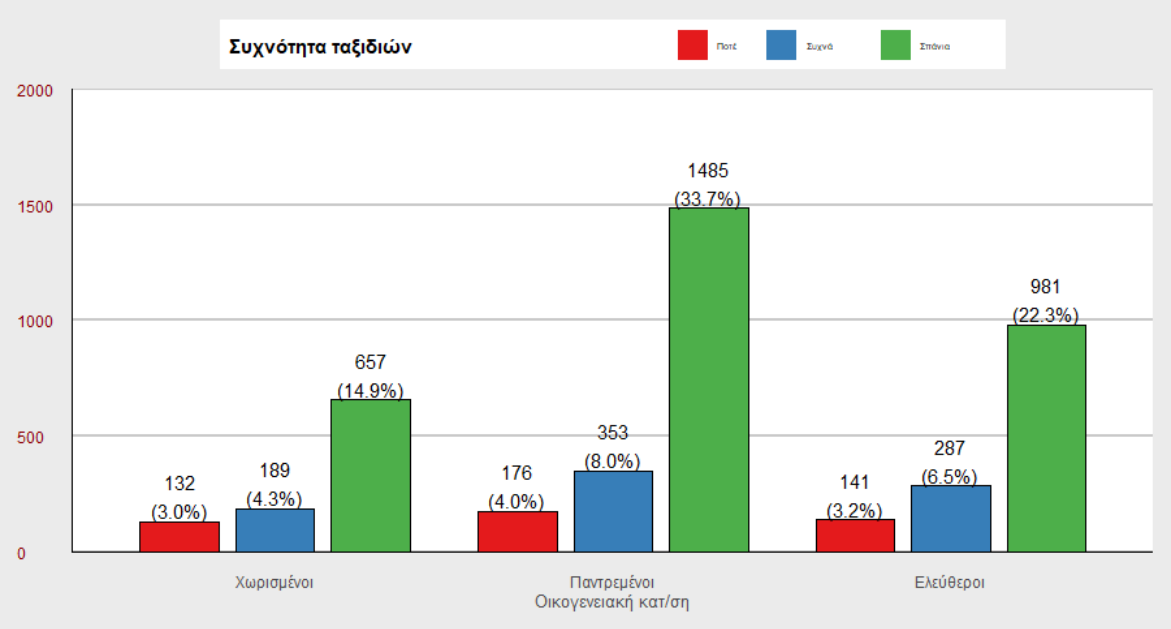
		Συνολική προϋπηρεσία (χρόνια)	
		Μέση τιμή	Τυπική απόκλιση
Συχνότητα ταξιδιών	Ποτέ	10.59243	7.418367
	Συχνά	11.09047	7.5380057
	Σπάνια	11.42907	7.892273

Πίνακας 5.3: *Συχνότητα ταξιδιών των υπαλλήλων ως προς την οικογενειακή κατάσταση / το τμήμα της εταιρείας.*

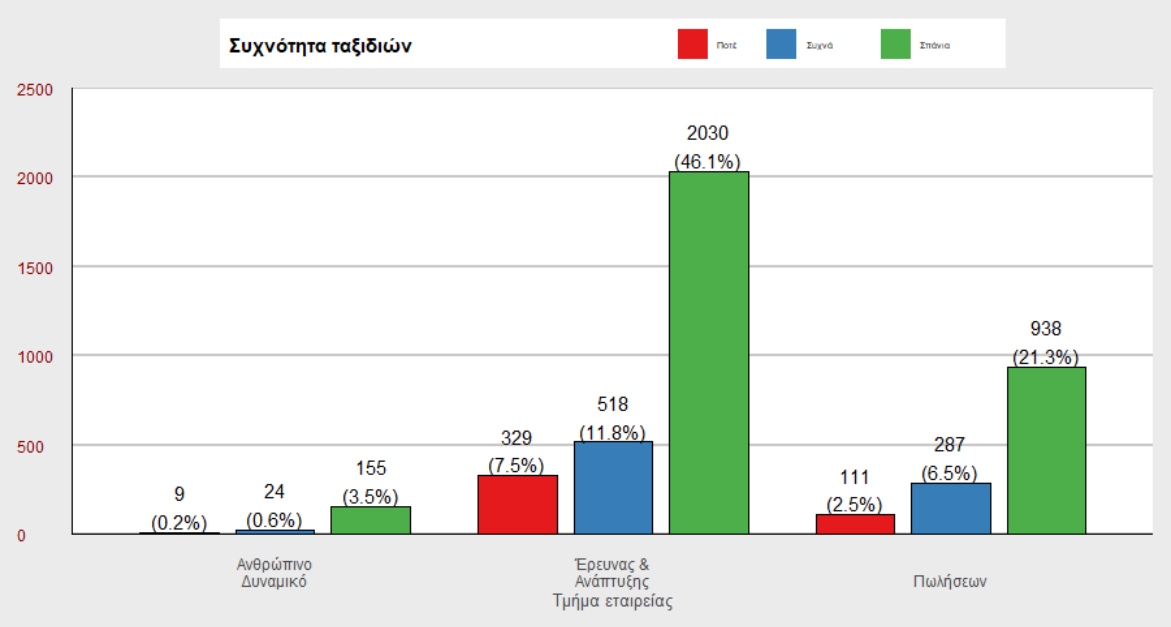
		Συχνότητα ταξιδιών			
		Ποτέ	Συχνά	Σπάνια	
Οικογενειακή κατ/ση	Χωρισμένος/η	132	189	657	978 / **(22.2%)**
	Παντρεμένος/η	176	353	1485	2014 / **(45.8%)**
	Ελεύθερος/η	141	287	981	1409 / **(32.0%)**
Τμήμα	Ανθρώπινου Δυναμικού	9	24	155	188 / **(4.3%)**
	Έρευνας & Ανάπτυξης	329	518	2030	2877 / **(65.4%)**
	Πωλήσεων	111	287	938	1336 / **(30.3%)**

Πίνακας 5.4: *Συχνότητα ταξιδιών των υπαλλήλων ως προς την οικογενειακή κατάσταση και το τμήμα της εταιρείας.*

Οικογενειακή κατ/ση	Τμήμα εταιρείας	Συχνότητα ταξιδιών			
		Ποτέ	Συχνά	Σπάνια	
Χωρισμένοι	Ανθρώπινο Δυναμικό	0	0	21	21 / **(0.5%)**
	Έρευνας & Ανάπτυξης	99	114	406	619 / **(14.0%)**
	Πωλήσεων	33	75	230	338 / **(7.6%)**
Παντρεμένοι	Ανθρώπινο Δυναμικό	3	18	74	95 / **(2.2%)**
	Έρευνας & Ανάπτυξης	122	225	1000	1347 / **(30.6%)**
	Πωλήσεων	51	110	411	572 / **(13.0%)**
Ελεύθεροι	Ανθρώπινο Δυναμικό	6	6	60	72 / **(1.6%)**
	Έρευνας & Ανάπτυξης	108	179	624	911 / **(20.8%)**
	Πωλήσεων	27	102	297	426 / **(9.7%)**
		449	829	3123	**4401 / (100%)**



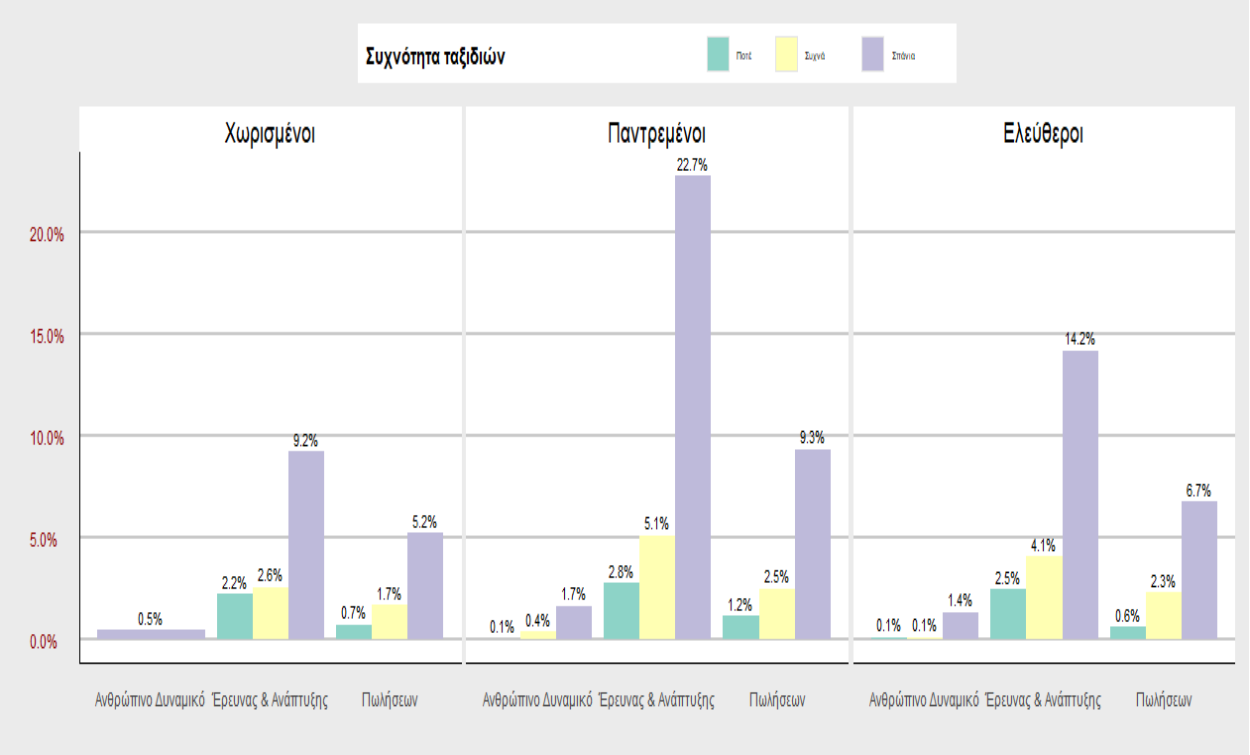
Σχήμα 5.1: *Συχνότητα ταξιδιών ως προς την οικογενειακή κατάσταση των υπαλλήλων.*



Σχήμα 5.2: *Συχνότητα ταξιδιών ως προς το τμήμα της εταιρείας που εργάζονται οι υπάλληλοι.*

Σχήμα 5.3: *Συνολικά έτη προϋπηρεσίας ως προς την συχνότητα των ταξιδιών και την οικογενειακή κατάσταση των υπαλλήλων.*

Σχήμα 5.4: *Συνολικά έτη προϋπηρεσίας ως προς την συχνότητα των ταξιδιών και το τμήμα της εταιρείας που εργάζονται οι υπάλληλοι.*



Σχήμα 5.5: *Συχνότητα ταξιδιών ως προς την οικογενειακή κατ/ση και το τμήμα της εταιρείας.*

5.2 Έλεγχος καλής προσαρμογής

Στην συνέχεια, προσαρμόζεται το πολυωνυμικό μοντέλο λογιστικής παλινδρόμησης με αποκρινόμενη την μεταβλητή Y=BusinessTravel και επεξηγηματικές τις X_1=MaritalStatus, X_2= Department και X_3= TotalWorkingYears. Ως κατηγορία αναφοράς (reference ή baseline category) θεωρούμε την κατηγορία ”Non-Travel” της μεταβλητής BusinessTravel. Στο σημείο αυτό είναι σημαντικό να τονισθεί, πως οι έλεγχοι καλής προσαρμογής αποτελούν σημαντικό μέρος της λογιστικής παλινδρόμησης, καθώς εαν αποδειχθεί ότι το μοντέλο δεν προσαρμόζεται καλά στα δεδομένα, οι ερμηνείες των αποτελεσμάτων της παλινδρόμησης καθίστανται λανθασμένες. Για τον λόγο αυτό, πρώτα θα ελεχθεί η καλή προσαρμογή του μοντέλου, σύμφωνα με τις μεθόδους που αναλύθηκαν παραπάνω για τον εντοπισμό του όρου αλληλεπίδρασης X_1X_3 που παραλείπεται και στην συνέχεια θα ακολουθήσουν συμπεράσματα της παλινδρόμησης.

Προσαρμόζουμε δυο μοντέλα, ένα με τον όρο αλληλεπίδρασης και ένα χωρίς,

$$\text{Μοντέλο I} \qquad ln\left(\frac{\pi_{ij}}{\pi_{i0}}\right) = b_{j0} + b_{j1}X_{i1} + b_{j2}X_{i2} + b_{j3}X_{i3}$$

$$\text{Μοντέλο II} \qquad ln\left(\frac{\pi_{ij}}{\pi_{i0}}\right) = b_{j0} + b_{j1}X_{i1} + b_{j2}X_{i2} + b_{j3}X_{i3} + b_{j4}X_{i1}X_{i3}$$

για $j = 1, 2$ όπου π_{ij} είναι η πιθάνοτητα το i-άτομο να ανήκει στην j-κατηγορία της αποκρινόμενης μεταβλητής και αντίστοιχα π_{i0} είναι η πιθάνοτητα το i-άτομο να ανήκει στην κατηγορία αναφοράς (”Ποτέ”). Γενικά, για να αντιμετωπισθεί το πρόβλημα του εντοπισμού ενός συγκεκριμένου όρου που παραλείπεται από το προσαρμοσμένο μοντέλο θα χρησιμοποιηθεί ο Έλεγχος Λόγου Πιθανοφανειών (Likelihood Ratio test). Ο έλεγχος LR χρησιμοποιείται για την σύγκριση των δυο μοντέλων, καθώς το Μοντέλο I είναι εμφωλευμένο (nested) στο Μοντέλο II. Έχει μεγάλη ισχύ και αποτελεί ένας καλός και εύκολος τρόπος να αποφασίσουμε μεταξύ των δυο μοντέλων, ποιο είναι το σωστό και κατά συνέπεια ποιοι από τους ελέγχους παίρνουν τις ορθές αποφάσεις για την προσαρμογή κάθε μοντέλου. Από τον Πίνακα 5.5 έπεται ότι ο όρος αλληλεπίδρασης είναι στατιστικά σημαντικός σε 5% ε.σ. και επομένως πρέπει να συμπεριληφθεί στην ανάλυση.

Πίνακας 5.5: *Έλεγχος λόγου πιθανοφανειών για την σύγκριση των δυο μοντέλων σε ε.σ. 5%.*

	Model	Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
1	Model I	8790	6902.851				
2	Model II	8786	6891.210	1 vs 2	4	11.6409	0.0202313

Για να ελέγξουμε την καταλληλότητα των ελέγχων καλής προσαρμογής που αναλύθηκαν στο Κεφάλαιο 4, αρκεί για κάθε μοντέλο να εφαρμόσουμε τους ελέγχους $\hat{C_{10}}$, X^2_{w*10}, X^2_{k*10}, X^2_{p*10} και να συγκρίνουμε τις αποφάσεις αυτών μεταξύ τους. Εφόσον ο όρος αλληλεπίδρασης X_1X_3 είναι στατιστικά σημαντικός, έπεται ότι το Μοντέλο II είναι καλύτερο από το Μοντέλο I. Συνεπώς, αναμένεται η απόρριψη του Μοντέλου I και η μη απόρριψη του Μοντέλου II από τους ελέγχους. Στον Πίνακα 5.6 δίνονται οι τιμές των τεσσάρων ελεγχοσυναρτήσεων με τα αντίστοιχα P-values για κάθε μοντέλο και στον Πίνακα 5.7 δίνονται αντίστοιχα οι αποφάσεις για κάθε μοντέλο.

Πίνακας 5.6: *Τιμές ελεγχοσυναρτήσεων και P-values για τα δυο μοντέλα σε 5% ε.σ.*

Έλεγχοι	Τιμές Ελεγχοσυναρτήσεων	$\chi^2_{16;0.05}$ [1]	P-values
Μοντέλο Ι			
$\hat{C_{10}}$	53.464	26.296	6.342e-06 < 0.05
X^2_{w*10}	24.679	26.296	0.076 > 0.05
X^2_{k*10}	30.199	26.296	0.017 < 0.05
X^2_{p*10}	27.255	26.296	0.039 < 0.05
Μοντέλο ΙΙ			
$\hat{C_{10}}$	30.732	26.296	0.014 < 0.05
X^2_{w*10}	13.969	26.296	0.601 > 0.05
X^2_{k*10}	19.298	26.296	0.254 > 0.05
X^2_{p*10}	24.567	26.296	0.078 > 0.05

Πίνακας 5.7: *Αποφάσεις των τεσσάρων ελέγχων για τα δυο μοντέλα σε 5% ε.σ.*

	Απόφαση για κάθε μοντέλο	
Έλεγχοι	**Μοντέλο Ι**	**Μοντέλο ΙΙ**
$\hat{C_{10}}$	Απόρριψη	Απόρριψη
X^2_{w*10}	Μη Απόρριψη	Μη Απόρριψη
X^2_{k*10}	Απόρριψη	Μη Απόρριψη
X^2_{p*10}	Απόρριψη	Μη Απόρριψη

5.2.1 Συμπεράσματα

Από τους παραπάνω πίνακες παρατηρείται πως μόνο ο X^2_{p*10} δεν απορρίπτει το σωστό μοντέλο που υποδεικνύει ο έλεγχος LR, δηλαδή το Μοντέλο ΙΙ, ενώ απορρίπτει ορθά το Μοντέλο Ι. Ωστόσο, αν και φαίνεται πως και ο X^2_{k*10} διεξάγει τις σωστές αποφάσεις για τα δυο μοντέλα, δεν θα ληφθεί υπόψιν, διότι ο k-means για διαφορετικά «τρεξίματα» καταλήγει πάντα σε διαφορετική ομαδοποίηση, καθώς κάθε φορά ο αλγόριθμος θεωρεί διαφορετικά

[1] $\chi^2_{df;0.05} = \chi^2_{16;0.05}$ όπου $df = (g-2) \times (c-1) = 16$

σημεία ως τα κέντρα της ομάδας (συστάδας). Στην συγκεκριμένη εφαρμογή, η διαφορετική ομαδοποίηση του αλγορίθμου οδηγεί κάθε φορά και σε διαφορετική απόφαση. Από την άλλη πλευρά, ο $\hat{C_{10}}$ αδυνατεί να μην απορρίψει το σωστό μοντέλο, ενώ ο X^2_{w*10} δεν βρίσκει ενδείξεις να απορρίψει το λάθος μοντέλο. Τα αποτελέσματα της εφαρμογής έρχονται σε συμφωνία με αυτά της προσομοίωσης, όσον αφορά την ισχύ του X^2_{p*10}. Παρόλ' αυτά, αναμέναμε μεγαλύτερη ισχύ από τον $\hat{C_{10}}$. Όπως αναφέρθηκε και στο Κεφάλαιο 3, η επιλογή του πλήθους των ομάδων καθώς και το μεγάλο μέγεθος δείγματος επηρεάζουν την ισχύ του $\hat{C}_g$. Σε κάθε περίπτωση, όπως με κάθε έλεγχο καλής προσαρμογής, ένα στατιστικά μη σημαντικό αποτέλεσμα δεν συνεπάγεται απαραίτητα ότι το μοντέλο έχει ικανοποιητική προσαρμογή στα δεδομένα. Θα πρέπει να γίνει περαιτέρω ανάλυση, όπως ο υπολογισμός του ποσοστού σωστής πρόβλεψης, του ψευδοσυντελεστή προσδιορισμού του μοντέλου [2] R^2, οι καμπύλες ROC για το δυαδικό λογιστικό μοντέλο, καθώς επίσης και έλεγχοι στατιστικής σημαντικότητας των συντελεστών παλινδρόμησης, ίσως να χρειάζονται για τον προσδιορισμό της προσαρμογής του μοντέλου (Xie et al.(2008) [19]).

Μια συνήθης προσέγγιση για την αξιολόγηση του λογιστικού μοντέλου παλινδρόμησης είναι ο υπολογισμός του *ποσοστού σωστής πρόβλεψης (ή ταξινόμησης)*. Για τον υπολογισμό αυτού, γίνεται διαχωρισμός του αρχικού συνόλου σε δυο υποσύνολα : το σύνολο εκπαίδευσης (train dataset) και το σύνολο ελέγχου (test dataset). Για το σύνολο εκπαίδευσης, θεωρούμε τυχαίο δείγμα, συνήθως το 70% του αρχικού συνόλου, και για το συγκεκριμένο σύνολο προσαρμόζεται το μοντέλο και υπολογίζονται οι αντίστοιχες εκτιμώμενες πιθανότητες και κατά συνέπεια οι κατηγορίες που προβλέπει το μοντέλο για κάθε άτομο. Η ίδια διαδικασία επαναλαμβάνεται και για το σύνολο ελέγχου. Έπειτα, δημιουργείται ένας πίνακας που απεικονίζει το πλήθος των ατόμων που ταξινομούνται σε κάθε κατηγορία, δεδομένης της κατηγορίας που ανήκουν. Στο παράδειγμα προκύπτει ένας 3×3 πίνακας, όπου στην κύρια διαγώνιο είναι το πλήθος των ατόμων που ταξινομούνται σωστά. Το ποσοστό σωστής ταξινόμησης του μοντέλου υπολογίζεται αθροίζοντας τα στοιχεία της κύριας διαγωνίου προς το σύνολο όλων των στοιχείων του πίνακα για καθένα από τα δυο σύνολα (train-test dataset). Στην συγκεκριμένη εφαρμογή λαμβάνονται ποσοστά σωστής ταξινόμησης 71.02% και 70.91% για κάθε υποσύνολο αντίστοιχα. Τα ποσοστά είναι υψηλά, γεγονός που οφείλεται στο πλήθος των ατόμων που ανήκουν στην κατηγορία Travel_Rarely της αποκρινόμενης μεταβλητής

[2] Δεν είναι αξιόπιστο και συνήθως δεν λαμβάνεται υπόψιν. Αποτελεί ένα μέτρο προσδιορισμού του μοντέλου σε αντιστοιχία με την γραμμική παλινδρόμηση. Οδηγεί συχνά σε πολύ χαμηλές τιμές.

BusinessTravel. Όπως είναι λογικό, το μοντέλο προβλέπει συνέχεια την κατηγορία αυτή, καθώς είναι πιο πιθανή να συμβεί μεταξύ των υπολοίπων, ενώ δεν προβλέπει ποτέ σωστά τις κατηγορίες Non_Travel και Travel_Frequently, αφού τις ταξινομεί λανθασμένα στην κατηγορία Travel_Rarely. Το γεγονός ότι προκύπτουν υψηλά ποσοστά σωστής πρόβλεψης δεν συναπάγεται απαραίτητα πως αυτά αντιστοιχούν και σε ορθές προβλέψεις. Ουσιαστικά, από τις 4401 παρατηρήσεις της αποκρινόμενης BusinessTravel, οι 3123 ανήκουν στην κατηγορία Travel_Rarely, οι οποίες ταξινομούνται ορθά στην σωστή κατηγορία, εξού και τα μεγάλα ποσοστά σωστής πρόβλεψης.

5.3 Ερμηνεία αποτελεσμάτων παλινδρόμησης

Αν και το μοντέλο δεν φαίνεται να είναι κατάλληλο για την πρόβλεψη της συχνότητας των ταξιδιών που κάνει ένας υπάλληλος για επαγγελματικούς σκοπούς, για λόγους κατανόησης, θα ερμηνεύσουμε ορισμένα από τα αποτελέσματα που προκύπτουν από την προσαρμογή του Μοντέλου ΙΙ στα δεδομένα. Τελικά για το προσαρμοσμένο Μοντέλο ΙΙ προκύπτουν τα παρακάτω αποτελέσματα.

- **Περίπτωση 1**

Θεωρώντας κατηγορίες αναφοράς για την αποκρινόμενη BusinessTravel την κατηγορία ”Non-Travel” και για τις επεξηγηματικές MaritalStatus και Department τις ”Divorced” και ”Human Resources” αντίστοιχα προκύπτουν δυο logit μετασχηματισμοί,

$$ln\left(\frac{P(BusinessTravel = Travel\_Frequently)}{P(BusinessTravel = Non\_Travel)}\right) = 0.4706361 + 0.1644996 * Married + 0.8199038 * Single -$$
$$-0.5015519 * Research\&Development + 0.0116278 * Sales +$$
$$+0.0184062 * TotalWorkingYears +$$
$$+0.0192073 * Married : TotalWorkingYears -$$
$$-0.0422113 * Single : TotalWorkingYears$$

$$ln\left(\frac{P(BusinessTravel = Travel\_Rarely)}{P(BusinessTravel = Non\_Travel)}\right) = 2.3077709 + 0.2883329 * Married + 0.5857833 * Single -$$
$$-1.0058122 * Research\&Development - 0.6606447 * Sales +$$
$$+0.0154851 * TotalWorkingYears +$$
$$+0.0229448 * Married : TotalWorkingYears -$$
$$-0.0230552 * Single : TotalWorkingYears$$

Πίνακας 5.8: *Travel_Frequently vs Non-Travel*

						95% CI για τα OR	
	Coefficient	Std. Errors	Z-stastic	p-value	OR	lower bound	upper bound
Intercept	0.4706361	0.4400622	1.0694763	0.2848551	1.6010124	0.7385062	2.4635185
Married	0.1644996	0.2660198	0.6183735	0.5363291	1.1788031	0.6574139	1.7001923
Single	0.8199038	0.2712629	3.0225433	**0.0025066**	2.2702815	1.7386160	2.8019469
Research & Development	-0.5015519	0.3980867	-1.2599063	0.2077032	0.6055901	-0.1746455	1.3858257
Sales	0.0116278	0.4076342	0.0285250	0.9772434	1.0116956	0.2127473	1.8106440
TotalWorkingYears	0.0184062	0.0151972	1.2111605	0.2258339	1.0185767	0.9887907	1.0483626
Married : TotalWorkingYears	0.0192073	0.0200834	0.9563793	0.3388806	1.0193930	0.9800303	1.0587557
Single : TotalWorkingYears	-0.0422113	0.0202996	-2.0794142	**0.0375793**	0.9586672	0.9188807	0.9984537

[a] Κατηγορία αναφοράς για την μεταβλητή MaritalStatus είναι η κατηγορία Divorced.

[b] Κατηγορία αναφοράς για την μεταβλητή Department είναι η κατηγορία Human Resources.

Πίνακας 5.9: *Travel_Rarely vs Non-Travel*

						95% CI για τα OR	
	Coefficient	Std. Errors	Z-stastic	p-value	OR	lower bound	upper bound
Intercept	2.3077709	0.3819886	6.041465	**0.0000000**	10.0519932	9.3033092	10.800677
Married	0.2883329	0.2252793	1.279891	0.2005835	1.3342014	0.8926621	1.775741
Single	0.5857833	0.2313904	2.531580	**0.0113550**	1.7963976	1.3428807	2.249915
Research & Development	-1.0058122	0.3488557	-2.883176	**0.0039369**	0.3657475	-0.3179971	1.049492
Sales	-0.6606447	0.3583145	-1.843757	0.0652186	0.5165182	-0.1857653	1.218802
TotalWorkingYears	0.0154851	0.0129709	1.193834	0.2325427	1.0156056	0.9901831	1.041028
Married : TotalWorkingYears	0.0229448	0.0175074	1.310574	0.1900017	1.0232100	0.9888961	1.057524
Single : TotalWorkingYears	-0.0230552	0.0172002	-1.340408	0.1801127	0.9772085	0.9434968	1.010920

[a] Κατηγορία αναφοράς για την μεταβλητή MaritalStatus είναι η κατηγορία Divorced.

[b] Κατηγορία αναφοράς για την μεταβλητή Department είναι η κατηγορία Human Resources.

- **Περίπτωση 2**

Θεωρώντας κατηγορίες αναφοράς για την αποκρινόμενη BusinessTravel την κατηγορία ”Non-Travel” και για τις επεξηγηματικές MaritalStatus και Department τις ”Married” και ”Sales” αντίστοιχα προκύπτουν δυο logit μετασχηματισμοί,

$$ln\left(\frac{P(BusinessTravel = Travel\_Frequently)}{P(BusinessTravel = Non\_Travel)}\right) = 0.6468563 - 0.1645131 * Divorced + 0.6554848 * Single -$$
$$-0.0115267 * HumanResources - 0.5132653 * Research\&Development$$
$$+0.0376120 * TotalWorkingYears -$$
$$-0.0192021 * Divorced : TotalWorkingYears -$$
$$-0.0614214 * Single : TotalWorkingYears$$

$$ln\left(\frac{P(BusinessTravel = Travel\_Rarely)}{P(BusinessTravel = Non\_Travel)}\right) = 1.9355331 - 0.2883420 * Divorced + 0.2974801 * Single +$$
$$+0.6608468 * HumanResources - 0.3452286 * Research\&Development +$$
$$+0.0384272 * TotalWorkingYears -$$
$$-0.0229379 * Divorced : TotalWorkingYears -$$
$$-0.0460003 * Single : TotalWorkingYears$$

Πίνακας 5.10: *Travel_Frequently vs Non-Travel*

						95% CI για τα OR	
	Coefficient	Std. Errors	Z-stastic	p-value	OR	lower bound	upper bound
Intercept	0.6468563	0.1905561	3.3945720	**0.0006874**	1.9095285	1.5360454	2.2830115
Divorced	-0.1645131	0.2660226	-0.6184177	0.5363001	0.8483067	0.3269120	1.3697014
Single	0.6554848	0.2402877	2.7279167	**0.0063736**	1.9260761	1.4551208	2.3970313
Human Resources	-0.0115267	0.4076706	-0.0282745	0.9774432	0.9885395	0.1895198	1.7875592
Research & Development	-0.5132653	0.1327631	-3.8660230	**0.0001106**	0.5985380	0.3383271	0.8587489
TotalWorkingYears	0.0376120	0.0131358	2.8633248	**0.0041922**	1.0383283	1.0125826	1.0640739
Divorced : TotalWorkingYears	-0.0192021	0.0200837	-0.9561059	0.3390187	0.9809811	0.9416179	1.0203443
Single : TotalWorkingYears	-0.0614214	0.0187984	-3.2673713	**0.0010855**	0.9404269	0.9035827	0.9772711

[a] Κατηγορία αναφοράς για την μεταβλητή MaritalStatus είναι η κατηγορία Married.

[b] Κατηγορία αναφοράς για την μεταβλητή Department είναι η κατηγορία Sales.

Πίνακας 5.11: *Travel_Rarely vs Non-Travel*

						95% CI για τα OR	
	Coefficient	Std. Errors	Z-stastic	p-value	OR	lower bound	upper bound
Intercept	1.9355331	0.1668659	11.599336	**0.0000000**	6.9277361	6.6006850	7.2547871
Divorced	-0.2883420	0.2252827	-1.279912	0.2005762	0.7495052	0.3079593	1.1910512
Single	0.2974801	0.2092773	1.421464	0.1551820	1.3464616	0.9362856	1.7566376
Human Resources	0.6608468	0.3583534	1.844120	0.0651656	1.9364314	1.2340716	2.6387912
Research & Development	-0.3452286	0.1173153	-2.942741	**0.0032532**	0.7080585	0.4781247	0.9379923
TotalWorkingYears	0.0384272	0.0117636	3.266626	**0.0010884**	1.0391751	1.0161189	1.0622313
Divorced : TotalWorkingYears	-0.0229379	0.0175077	-1.310158	0.1901423	0.9773232	0.9430087	1.0116377
Single : TotalWorkingYears	-0.0460003	0.0163005	-2.822024	**0.0047722**	0.9550416	0.9230933	0.9869900

[a] Κατηγορία αναφοράς για την μεταβλητή MaritalStatus είναι η κατηγορία Married.

[b] Κατηγορία αναφοράς για την μεταβλητή Department είναι η κατηγορία Sales.

Πίνακας 5.12: *Ψευδο R^2 και AIC με κατηγορία αναφοράς την "Non-Travel"*

	Ψευδοσυντελεσής R^2			
	CoxSnell	McFadden	Nagelkerke	AIC
Περίπτωση 1	0.015546256	0.009907303	0.019571461	6923.21
Περίπτωση 2	0.015546256	0.009907303	0.019571461	6923.21

- Τέλος, για να λάβουμε συγκρίσεις μεταξύ των υπαλλήλων που ταξιδεύουν σπάνια σε σχέση με εκείνους που ταξιδεύουν συχνά, επαναλαμβάνεται η παραπάνω διαδικασία. Ως κατηγορία αναφοράς για την αποκρινόμενη BusinessTravel θεωρείται η κατηγορία "Travel_Frequently" και για τις επεξηγηματικές MaritalStatus και Department θεωρούνται αντίστοιχα τα ζεύγη κατηγοριών "Divorced" - "Human Resources" και "Married" - "Sales" ως κατηγορίες αναφοράς, όπως φαίνεται από τους Πίνακες 5.13 και 5.14. Επιπλέον, οι πίνακες με τις συγκρίσεις Non-Travel vs Travel_Frequently παραλείπονται και για τις δυο περιπτώσεις, καθώς είναι ίδιοι με τους Πίνακες 5.8 και 5.10 με αντίθετα πρόσημα.

Πίνακας 5.13: *Travel_Rarely vs Travel_Frequently*

						95% CI για τα OR	
	Coefficient	Std. Errors	Z-stastic	p-value	OR	lower bound	upper bound
Intercept	1.8371997	0.2653024	6.9249262	**0.0000000**	6.2789307	5.7589475	6.7989139
Married	0.1239027	0.1862535	0.6652367	0.5058991	1.1319057	0.7668555	1.4969559
Single	-0.2340505	0.1899030	-1.2324741	0.2177720	0.7913218	0.4191188	1.1635249
Research & Development	-0.5043484	0.2254320	-2.2372532	**0.0252698**	0.6038989	0.1620603	1.0457375
Sales	-0.6723677	0.2301886	-2.9209433	**0.0034897**	0.5104984	0.0593371	0.9616597
TotalWorkingYears	-0.0029219	0.0106789	-0.2736134	0.7843818	0.9970824	0.9761522	1.0180126
Married:TotalWorkingYears	0.0037341	0.0130172	0.2868572	0.7742217	1.0037411	0.9782278	1.0292543
Single:TotalWorkingYears	0.0191515	0.0142023	1.3484782	0.1775046	1.0193360	0.9915001	1.0471720

[a] Κατηγορία αναφοράς για την μεταβλητή MaritalStatus είναι η κατηγορία Divorced.

[b] Κατηγορία αναφοράς για την μεταβλητή Department είναι η κατηγορία Human Resources.

Πίνακας 5.14: *Travel_Rarely vs Travel_Frequently*

						95% CI για τα OR	
	Coefficient	Std. Errors	Z-stastic	p-value	OR	lower bound	upper bound
Intercept	1.2886731	0.1198699	10.7505952	**0.0000000**	3.6279694	3.3930287	3.862910
Divorced	-0.1237833	0.1862551	-0.6645904	0.5063125	0.8835713	0.5185181	1.248625
Single	-0.3579709	0.1551456	-2.3073218	**0.0210369**	0.6990934	0.3950136	1.003173
Human Resources	0.6723601	0.2301857	2.9209468	**0.0034897**	1.9588551	1.5076994	2.410011
Research & Development	0.1680195	0.0837913	2.0052132	**0.0449403**	1.1829597	1.0187317	1.347188
TotalWorkingYears	0.0008162	0.0074493	0.1095675	0.9127524	1.0008165	0.9862162	1.015417
Divorced : TotalWorkingYears	-0.0037385	0.0130174	-0.2871935	0.7739641	0.9962685	0.9707549	1.021782
Single : TotalWorkingYears	0.0154198	0.0119639	1.2888593	0.1974470	1.0155393	0.9920905	1.038988

[a] Κατηγορία αναφοράς για την μεταβλητή MaritalStatus είναι η κατηγορία Married.

[b] Κατηγορία αναφοράς για την μεταβλητή Department είναι η κατηγορία Sales.

Πίνακας 5.15: *Ψευδο R^2 και AIC με κατηγορία αναφοράς την "Travel_Frequently"*

	Ψευδοσυντελεσής R^2			
	CoxSnell	McFadden	Nagelkerke	AIC
Περίπτωση 1	0.015546255	0.009907303	0.019571460	6923.21
Περίπτωση 2	0.015546256	0.009907303	0.019571461	6923.21

Με βάση τους παραπάνω πίνακες, θα ακολουθήσει σχολιασμός μόνο των στατιστικά σημαντικών συντελεστών, με στόχο την κατανόηση της ερμηνείας αυτών για το πολωνυμικό λογιστικό μοντέλο παλινδρόμησης.

- **Σε σχέση με την οικογενειακή κατάσταση του υπαλλήλου** :

Τα odds ενός υπαλλήλου της εταιρείας να κάνει *συχνά* ταξίδια για επαγγελματικούς σκοπούς από το να μην κάνει *ποτέ*, είναι $2.27 - 1 = 1.27$ φορές μεγαλύτερα για τους ελεύθερους συγκριτικά με τους διαζευγμένους υπαλλήλους και 93% [ή $(1.93 - 1) * 100\%$] μεγαλύτερα σε σχέση με τους παντρεμένους υπαλλήλους, ανεξάρτητα από το τμήμα της εταιρείας και τα συνολικά έτη προϋπηρεσίας.

Τα odds ενός υπαλλήλου της εταιρείας να ταξιδεύει *σπάνια* από το να μην ταξιδεύει *ποτέ* για επαγγελματικούς σκοπούς, είναι 80% [ή $(1.80 - 1) * 100\%$] μεγαλύτερα για τους ελεύθερους υπαλλήλους σε σχέση με τους διαζευγμένους , ανεξάρτητα από το τμήμα της εταιρείας και τα συνολικά έτη προϋπηρεσίας.

Τέλος, τα odds ενός υπαλλήλου της εταιρείας να ταξιδεύει *σπάνια* για επαγγελματικούς σκοπούς από το να ταξιδεύει *συχνά*, είναι χαμηλότερα κατά 30% [ή $(1 - 0.70) * 100\%$] για τους ελεύθερους συγκριτικά με τους παντρεμένους υπαλλήλους, ανεξάρτητα από το τμήμα της εταιρείας και τα συνολικά έτη προϋπηρεσίας.

- **Σε σχέση με το τμήμα της εταιρείας που εργάζεται ο υπάλληλος :**

Τα odds ενός υπαλλήλου να ταξιδεύει *συχνά* από το να μην ταξιδεύει *ποτέ* για επαγγελματικούς σκοπούς, είναι 40% [ή $(1 - 0.60) * 100\%$] χαμηλότερα για τους υπαλλήλους που εργάζονται στο Τμήμα Έρευνας και Ανάπτυξης σε σχέση με εκείνους που εργάζονται στο Τμήμα Πωλήσεων της εταιρείας, ανεξάρτητα από την οικογενειακή κατάσταση και τα συνολικά έτη προϋπηρεσίας.

Τα odds ενός υπαλλήλου να ταξιδεύει *σπάνια* από το να μην ταξιδεύει *ποτέ* για επαγγελματικούς σκοπούς, είναι 63% [ή $(1 - 0.37) * 100\%$] χαμηλότερα για τους υπαλλήλους που εργάζονται στο Τμήμα Έρευνας και Ανάπτυξης σε σχέση με εκείνους που εργάζονται στο Τμήμα Ανθρώπινου Δυναμικού και αντίστοιχα 29% [ή $(1 - 0.71) * 100\%$] χαμηλότερα σε σχέση με τα άτομα που εργάζονται στο Τμήμα Πωλήσεων, ανεξάρτητα από την οικογενειακή κατάσταση και τα συνολικά έτη προϋπηρεσίας.

Τα odds ενός υπαλλήλου να ταξιδεύει *σπάνια* από το να ταξιδεύει *συχά* για επαγγελματικούς σκοπούς, είναι 40% [ή $(1 - 0.60) * 100\%$] χαμηλότερα για τους υπαλλήλους που εργάζονται στο Τμήμα Έρευνας και Ανάπτυξης σε σχέση με εκείνους που εργάζονται στο Τμήμα Ανθρώπινου Δυναμικού, 18% [ή $(1.18 - 1) * 100\%$] μεγαλύτερα σε σχέση με τα άτομα που εργάζονται στο Τμήμα Πωλήσεων και 96% [ή $(1.96 - 1) * 100\%$] φορές μεγαλύτερα για τους υπαλλήλους που εργάζονται στο Τμήμα Ανθρώπινου Δυναμικού σε σχέση με εκείνους που εργάζονται στο Τμήμα Πωλήσεων της εταιρείας, ανεξάρτητα από την οικογενειακή κατάσταση και τα συνολικά έτη προϋπηρεσίας.

- **Σε σχέση με τα συνολικά χρόνια προϋπηρεσίας του υπαλλήλου :**

Τα odds ενός υπαλλήλου της εταιρείας να ταξιδεύει *συχνά (αντίστοιχα σπάνια)* από το να μην ταξιδεύει *ποτέ* για επαγγελματικούς σκοπούς, αυξάνουν κατά 4% [ή $(1.04 - 1) * 100\%$] για κάθε επιπλέον έτος προϋπηρεσίας του υπαλλήλου, ανεξάρτητα από την οικογενειακή κατάσταση και το τμήμα της εταιρείας που εργάζεται.

- **Σε σχέση με την αλληλεπίδραση της οικογενειακής κατάστασης και της συνολικής προϋπηρεσίας :**

Τα odds ενός υπαλλήλου της εταιρείας να ταξιδεύει *συχνά* από το να μην ταξιδεύει *ποτέ* για επαγγελματικούς σκοπούς, μειώνονται κατά 4% [ή $(1 - 0.96) * 100\%$] για κάθε επιπλέον έτος προϋπηρεσίας ενός ελεύθερου υπαλλήλου συγκριτικά με ενός διαζευγμένου. Αντίστοιχα, τα odds μειώνονται κατά 6% [ή $(1 - 0.94) * 100\%$] για κάθε επιπλέον έτος προϋπηρεσίας ενός ελεύθερου υπαλλήλου σε σχέση με έναν παντρεμένο υπάλληλο.

Τα odds ενός υπαλλήλου της εταιρείας να ταξιδεύει *σπάνια* από το να μην ταξιδεύει *ποτέ* για επαγγελματικούς σκοπούς, μειώνονται κατά 4% [ή $(1 - 0.96) * 100\%$] για κάθε επιπλέον έτος προϋπηρεσίας ενός ελεύθερου συγκριτικά με ενός παντρεμένου υπαλλήλου.

- **Ερμηνεία των στατιστικά σημαντικών σταθερών όρων (intercepts):**

Ένας υπάλληλος, ο οποίος είναι παντρεμένος, εργάζεται στο Τμήμα Πωλήσεων της εταιρείας και έχει μηδενική προϋπηρεσία, είναι 91% [ή $(1.91 - 1) * 100\%$] πιθανότερο να ταξιδεύει *συχνά* από το να μην ταξιδεύει *ποτέ* για επαγγελματικούς σκοπούς, $6.9 - 1 = 5.9$ φορές πιθανότερο να ταξιδεύει *σπάνια* από το να μην ταξιδεύει *ποτέ* για επαγγελματικούς

σκοπούς και $3.6 - 1 = 2.6$ φορές πιθανότερο να ταξιδεύει *σπάνια* από το να ταξιδεύει *συχνά* για επαγγελματικούς σκοπούς.

Ένας υπάλληλος, ο οποίος είναι χωρισμένος, εργάζεται στο Τμήμα Ανθρώπινου Δυναμικού της εταιρείας και έχει μηδενική προϋπηρεσία, είναι $10.1 - 1 = 9.1$ φορές πιθανότερο να ταξιδεύει *σπάνια* από το να μην ταξιδεύει *ποτέ* για επαγγελματικούς σκοπούς και $6.3 - 1 = 5.3$ φορές πιθανότερο να ταξιδεύει *σπάνια* από το να ταξιδεύει *συχνά* για επαγγελματικούς σκοπούς.

Κεφάλαιο 6

Συμπεράσματα και Προβληματισμοί

Στόχος της συγκεκριμένης εργασίας ήταν η παρουσίαση του λογιστικού μοντέλου και των ελέγχων καλής προσαρμογής, με έμφαση στο πολυωνυμικό μοντέλο λογιστικής παλινδρόμησης. Η νέα στρατηγική προτείνει την ομαδοποίηση των επεξηγηματικών μεταβλητών του μοντέλου, σύμφωνα με γνωστές μεθόδους συσταδοποίησης, Στην εργασία αυτή χρησιμοποιήθηκαν από τις ιεραρχικές μεθόδους, η μέθοδος του Ward και από τις διαχωριστικές, οι k-means και PAM. Ως μετρική απόστασης χρησιμοποιήθηκε η Ευκλείδεια. Ίσως, η χρήση διαφορετικών μετρικών απόστασης και διαφορετικών μεθόδων συσταδοποίησης να βελτιώσουν την ισχύ του προτεινόμενου ελέγχου [9]. Επιπλέον, η ισχύς των ελέγχων υπολογίσθηκε με βάση την ικανότητα τους να εντοπίσουν έναν όρο αλληλεπίδρασης ή έναν τετραγωνικό όρο που παραλείπεται από το πραγματικό μοντέλο. Ένα στατιστικά μη σημαντικό αποτέλεσμα, δεν συνεπάγεται ότι το μοντέλο προσαρμόζεται καλά στα δεδομένα. Περαιτέρω ανάλυση θα πρέπει να πραγματοποιηθεί για την απόφαση της καταλληλότητας του μοντέλου. Από την εφαρμογή, προέκυψαν ορισμένα συμπεράσματα και προβληματισμοί, οι οποίοι εαν ληφθούν υπόψιν είναι πιθανόν να βελτιώσουν την ισχύ του προτεινόμενου ελέγχου [19].

- Αρχικά, διακρίνεται κάποιου είδους μεροληψία για την πρόβλεψη των κατηγοριών της αποκρινόμενης BusinessTravel, εξαιτίας του άνισου πλήθους ατόμων σε κάθε κατηγορία. Το 71% των παρατηρήσεων της αποκρινόμενης αντιστοιχεί στους υπαλλήλους που ταξιδεύουν σπάνια (Travel_Rarely), με αποτέλεσμα το προσαρμοσμένο μοντέλο να προβλέπει συνέχεια αυτή την κατηγορία για κάθε περίπτωση. Τις μικρότερες κατηγορίες Non_Travel και Travel_Frequently αδυνατεί να τις προβλέψει, καθώς είναι λιγότερο πιθανές να συμβούν. Το γεγονός ότι υπάρχει μεροληψία στις κατηγορίες της

αποκρινόμενης, ίσως επηρεάζει τις αποφάσεις των ελέγχων για την συνολική αξιολόγηση του μοντέλου.

- Η συνεχής επεξηγηματική μεταβλητή TotalWorkingYears παρουσιάζει υψηλή ασυμμετρία. Παρόλ' αυτά, δεν φαίνεται να επηρεάζει την ισχύ του ελέγχου X^2_{p*10}.

- Ακόμη, παρατηρήθηκε πως διαφορετικές επιλογές της κατηγορίας αναφοράς οδήγησε σε διαφορετικές αποφάσεις των ελέγχων καλής προσαρμογής. Η επιλογή της κατηγορίας "Non-Travel" ως κατηγορίας αναφοράς ήταν τυχαία.

- Επιπλέον, η κωδικοποίηση των κατηγορικών επεξηγηματικών μεταβλητών που χρησιμοποιούνται για συσταδοποίηση παίζει πολύ σημαντικό ρόλο [18]. Διαφορετική κωδικοποίηση οδήγησε σε διαφορετικά αποτελέσματα. Για παράδειγμα, όταν οι κατηγορικές μεταβλητές αντιμετωπίζονται ως συνεχείς (αριθμητικές), τότε συνήθως χάνεται σημαντικό μέρος πληροφορίας. Στην εφαρμογή, δημιουργούνται ψευδομεταβλητές (dummy variables) για τις κατηγορικές επεξηγηματικές μεταβλητές, δηλαδή για κάθε κατηγορία δημιουργείται μια στήλη και οι παρατηρούμενες κατηγορίες κωδικοποιούνται με 1, ενώ οι υπόλοιπες με 0. Στην συνέχεια, δημιουργείται στην R ένα dataframe το οποίο έχει ως στήλες τις κατηγορίες των κατηγορικών επεξηγηματικών μεταβλητών και την μοναδική συνεχή μεταβλητή, εκτός από την αποκρινόμενη. Στην συνέχεια γίνεται τυποποίηση σε κάθε στήλη και τέλος πραγματοποιείται η διαδικασία της συσταδοποίησης με κάποια από τις τρεις μεθόδους.

- Τέλος, ο καθορισμός του πλήθος των συστάδων χρήζει προσοχής, διότι στην συγκεκριμένη εφαρμογή η μεταβολή του πλήθους αυτών οδήγησε σε διαφορετικές αποφάσεις.

Παράρτημα Α΄

Α΄.1 Σχετική πιθανότητα (Odds)

Δεδομένου ότι η λογιστική παλινδρόμηση προβλέπει την πιθανότητα εμφάνισης ενός γεγονότος, η επίδραση των επεξηγηματικών μεταβλητών πάνω σε αυτή εκφράζεται με την βοήθεια της *σχετικής πιθανότητας (odds)* [24]. Έστω π να είναι η πιθανότητα να συμβεί ένα γεγονός και $1-\pi$ αντίθετα. Τα odds ενός γεγονότος είναι ο λόγος των συμπληρωματικών πιθανοτήτων

$$odds\, of\, \{Event\} = \frac{\pi}{1-\pi} = \frac{\textit{πιθανότητα να συμβεί το γεγονός}}{\textit{πιθανότητα να μην συμβεί το γεγονός}}.$$

Τα odds υπερ της εμφάνισης ενός γεγονότος εκφράζονται με την μορφή $\frac{\pi}{1-\pi} : 1$.

Η σχετική πιθανότητα αποτελεί 1-1 μετασχηματισμός της πιθανότητας π

$$odds = \frac{\pi}{1-\pi} \Leftrightarrow \pi = \frac{odds}{odds+1}$$

που διευκολύνει την ερμηνεία των αποτελεσμάτων της παλινδρόμησης, καθώς συγκρίνει την πιθανότητα εμφάνισης ενός γεγονότος με την πιθανότητα μη εμφάνισης αυτού και όχι τις πιθανότητες αυτές καθαυτές.

Α΄.2 Λόγος σχετικών πιθανοτήτων (Odds Ratio,OR)

Ο λόγος σχετικών πιθανοτήτων (odds ratio) αποτελεί επέκταση του ορισμού της σχετικής πιθανότητας (odds) και ουσιαστικά συγκρίνει τα odds για διαφορετικές τιμές ή επίπεδα μιας δεύτερης μεταβλητής X. Έτσι ο λόγος σχετικών πιθανοτήτων του $X = 1$ έναντι του $X = 2$ [24] ισούται με

$$OR_{12} = \frac{odds(X = 1)}{odds(X = 2)} = \alpha, \tag{Α΄.1}$$

όπου

$$odds(X = x) = \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)}. \tag{Α΄.2}$$

Περιπτώσεις

- Αν $\alpha = 1$, τότε ισχύει $odds(X = 1) = odds(X = 2)$.
- Αν $\alpha > 1$, τότε ισχύει $odds(X = 1) > odds(X = 2)$.
- Αν $\alpha < 1$, τότε ισχύει $odds(X = 1) < odds(X = 2)$.

Ερμηνεία

- Όταν $OR_{12} = \alpha$, τότε τα odds του $Y = 1$ όταν $X = 1$ ισούται με α φορές τα αντίστοιχα odds όταν $X = 2$.
- Όταν $OR_{12} = \alpha > 1$, τότε τα odds του $Y = 1$ όταν $X = 1$ είναι $\alpha - 1$ φορές (ή $\{\alpha - 1\} * 100\%$) μεγαλύτερα από τα αντίστοιχα odds όταν $X = 2$.
- Όταν $OR_{12} = \alpha < 1$, τότε τα odds του $Y = 1$ όταν $X = 1$ είναι $1 - \alpha$ φορές (ή $\{1 - \alpha\} * 100\%$) μικρότερα από τα αντίστοιχα odds όταν $X = 2$.

Παράρτημα Β΄

Κώδικας εφαρμογής

```
library(openxlsx)
library(NbClust)
library(ResourceSelection)
library(moments)
library(MASS)
library(Rmisc)
library(fastcluster)
library(epiR)
library(nnet)
library(stats)
library(cluster)
library(generalhoslem)
library(cluster)
library(factoextra)
library(caret)
library(fpc)
library(e1071)
library(mlr)
library(DescTools)
library(car)
library(formattable)
library(knitr)
library(ggplot2)
library(ggpubr)
library(dplyr)
library(backports)
```

```
library(CGPfunctions)
library(ggpubr)
library(sjPlot)
library(sjmisc)
library(cowplot)
library(perturb)
library(ggthemes)
library(kableExtra)
library(effects)
library(broom)

data<-read.xlsx("C:/Users/data1.xlsx",skipEmptyRows = TRUE,skipEmptyCols
    = TRUE)
data1<-data[,c(3,4,10,12)]
data1<-data1[complete.cases(data1),]

data1$BusinessTravel<-factor(data1$BusinessTravel,levels = c("Non-Travel"
    ,"Travel\_Frequently","Travel\_Rarely"))
data1$Department<-factor(data1$Department,levels = c("Human Resources","
    Research \& Development","Sales" ))
data1$MaritalStatus<-factor(data1$MaritalStatus,levels = c("Divorced", "
    Married","Single"))
outcome<-data1$BusinessTravel
```

Περιγραφικά στατιστικά

```
#descriptives
summary(data1)

#for the categorical predictors

#BusinessTravel by MaritalStatus/Department
table(outcome)
table(data1$MaritalStatus,outcome)
tab1<-with(data1,table(data1$MaritalStatus,outcome))
tab2<-with(data1,table(data1$Department,outcome))
tab<- rbind(tab1,tab2)
knitr::kable(tab,format="latex")
```

```
13
14 #BusinessTravel by MaritalStatus and Department
15 tab3<-ftable(xtabs(~data1$MaritalStatus+data1$Department+data1$
       BusinessTravel,data=data1))
16 kable(tab3, "latex", booktabs = T) %>%
17 kable_styling(latex_options = "striped")
18
19 #for the continuous predictors
20 k<-with(data1, do.call(rbind, tapply(TotalWorkingYears, outcome, function
       (x) c(M = mean(x), SD = sd(x)))))
21 kable(k, "latex", booktabs = T) %>%
22 kable_styling(latex_options = "striped")
23
24 attach(data1)
25 #plotbars #+geom_color()/geom_size
26 set_theme(
27 base = theme_economist_white(),
28 axis.linecolor = "black",     # "remove" axis lines
29 axis.textcolor.y = "darkred", # set axis label text only for y axis
30 axis.tickslen = 0,            # "remove" tick marks
31 #legend.title.color = "magenta",   # legend title color
32 #legend.title.size = 2,         # legend title size
33 legend.color = "black",       # legend label color
34 legend.pos = "top",           # legend position above plot
35 axis.title.size = .9,
36 axis.textsize = .9,
37 legend.size = .5,
38 geom.label.size = 4
39 )
40 a <- ggplot(data1, aes(x=BusinessTravel, y=TotalWorkingYears,fill=
       MaritalStatus)) +geom_boxplot()
41 plot1<- a+scale_fill_brewer(palette="Pastel1")+labs(x="Travel Frequency",
       y=" TotalWorkingYears",fill="Marital Status")
42
43 b <- ggplot(data1, aes(x=BusinessTravel, y=TotalWorkingYears,fill=
       Department)) +geom_boxplot()
44 plot2<- b+scale_fill_brewer(palette="Pastel2")+labs(x="Travel Frequency",
       y="TotalWorkingYears",fill="Department")
```

```
#ggarrange(plot1,plot2,ncol = 2, nrow = 1)

plot3<-plot_grpfrq(MaritalStatus,BusinessTravel,legend.title = "Travel
    Frequency",axis.titles = "Marital Status",geom.colors = "Set1")
plot4<-plot_grpfrq(Department,BusinessTravel,legend.title = "Travel
    Frequency",axis.titles = "Department",geom.colors = "Set1")
#ggarrange(plot3,plot4,ncol =1, nrow = 2)

data1 %>%
count(cyl = BusinessTravel, MaritalStatus,Department) %>%
mutate(pct = prop.table(n)) %>%
ggplot(aes(x =Department , y = pct, fill = cyl, label = scales::percent(
    pct))) +
facet_wrap(~MaritalStatus)+geom_col(position = 'dodge') +
geom_text(position = position_dodge(width = .9), vjust = -0.5, size = 3)
    +
scale_y_continuous(labels = scales::percent)+scale_fill_brewer(palette="
    Set3")
```

'Ελεγχος καλής προσαρμογής με την χρήση μεθόδων συσταδοποίησης

```
#covariates
x1<-data1$MaritalStatus
x2<-data1$Department
x3<-data1$TotalWorkingYears

outcome<-data1$BusinessTravel
observed<-createDummyFeatures(outcome) # to create dummy variable for
    outcome variable
n_outcome<-ncol(observed)
outcome<-relevel(outcome,ref = "Non-Travel") #set reference/baseline
    category

# MODEL 1 : multinomial logistic regression WITHOUT interaction term (
    nested model)
model1<-multinom(outcome ~ x1 + x2 + x3 ,family="binomial",data = data1)
summary(model1)
#x1 + x2 + x3 + x1:x2 + x1:x3 best model from stepAIC
```

```
# MODEL 2 : multinomial logistic regression WITH interaction term (full model)
model2<-multinom(outcome ~ x1 + x2 + x3 + x1:x3,family="binomial",data = data1)
summary(model2)

#comparison of the 2 models : 1vs2 (without VS with)
anova(model1,model2,test="Chisq")
kable(anova(model1,model2,test="Chisq"), "latex", booktabs = T)

model<-model1

#observed values
for (i in 1:n_outcome) {
assign(paste0("observed", i-1),as.matrix(observed[,i]))
}

#expected values
expected<-fitted(model)
for (i in 1:n_outcome) {
assign(paste0("expected", i-1),as.matrix(expected[,i]))
}

#prepare covariates for clustering

#without outcome variable
new_df<-data1[,-1]

#recode categorical variables for cluster analysis

department<-createDummyFeatures(new_df$Department)
maritalStatus<-createDummyFeatures(new_df$MaritalStatus)
new_df<-data.frame(department,maritalStatus,new_df$TotalWorkingYears)
names(new_df)<-paste(c("dep.HR","dep.R&D","dep.Sales","divorced","married","single","TotalWorkingYears"))
new_df<-scale(new_df)

g  <-10   #number of groups/clusters
```

```
df <-16   #(g-2)(c-1)=degrees of freedom
a  <-0.05 #nominal level

#Distance metrics
d <- dist(new_df, method = "euclidean") # distance for numeric data types
   /recode categorical variables
#d<-daisy(new_df, metric = c("gower"))    # distance for mixed data types

#Hierarhical clustering method / ward's method

#hc <- hclust(d, method = "ward.D2")
#plot(hc) # display dendogram
#groups <- cutree(hc, g) # cut tree into k clusters
#rect.hclust(hc, g, border = "green") # draw dendogram with red borders
   around the k clusters

#Partitioning clustering methods

#PAM
pam<-pam(d, k=g)
groups<-pam$clustering

#K-means
#km.res <- kmeans(new_df,g,nstart = 25)
#groups <- km.res$cluster

new_data<-data.frame(outcome,x1,x2,x3,expected0,expected1,expected2,
   observed0,observed1,observed2,groups)
n<-nrow(data1)

new_data<-split(new_data,new_data$groups)

#observed values
for (i in 1:g) {
assign(paste0("O", i),data.frame(sum(new_data[[i]][8]), sum(new_data[[i
   ]][9]),sum(new_data[[i]][10])))
}

```

```
#expected values
for (i in 1:g) {
assign(paste0("E", i),data.frame(sum(new_data[[i]][5]), sum(new_data[[i
    ]][6]),sum(new_data[[i]][7])))
}

#number of rows for each cluster
for (i in 1:g) {
assign(paste0("ng", i), nrow(new_data[[i]]))
}

#Proposed test

for (i in 1:g) {
assign(paste0("mp", i), sum(((get(eval(paste0("O",i)))-get(eval(paste0("E
    ",i))))^2)/get(eval(paste0("E",i)))))
}

vmp<-rep(NA,g)
for (i in 1:g) {
vmp[i]<-sum(get(eval(paste0("mp",i))))
}

modified.pearson<-sum(vmp)

if(modified.pearson > qchisq(1-a, df=df)) #reject H0 : model fits the
    data/nominal level=5%
{
print(paste("H0 rejected"))
}else{
print(paste("H0 not rejected"))
}

#Cg based on grouping of predicted probabilities

Cg<-logitgof(outcome,fitted(model),g=g)

if(Cg$statistic > qchisq(1-a, df=df)) #reject H0
```

```
{
print(paste("H0 rejected"))
}else{
print(paste("H0 not rejected"))
}

pvalue_mod.pearson<-pchisq(modified.pearson, df=df, lower.tail=FALSE)
pvalue_Cg<-pchisq(Cg$statistic, df=df, lower.tail=FALSE)
print(paste("modified.pearson=", modified.pearson))
print(paste("Cg=",Cg$statistic))

print(paste("pvalue_modified.pearson=", pvalue_mod.pearson))
print(paste("pvalue_Cg=",pvalue_Cg))
```

Προσαρμογή και ανάλυση του πολυωνυμικού λογιστικού μοντέλου παλινδρόμησης

```
#train <- sample_frac(data1, 0.7)
#sample_id <- as.numeric(rownames(train))
#test <- data1[-sample_id,]

data1$BusinessTravel<-factor(data1$BusinessTravel,levels = c("Non-Travel"
    ,"Travel_Frequently","Travel_Rarely"))
data1$Department<-factor(data1$Department,levels = c("Human Resources","
    Research & Development","Sales" ))
data1$MaritalStatus<-factor(data1$MaritalStatus,levels = c("Divorced", "
    Married","Single"))

#reference categories/ Default : Non-Travel / Divorced / Human Resources
data1$BusinessTravel<-relevel(data1$BusinessTravel,ref = "Non-Travel")
#data1$MaritalStatus<-relevel(data1$MaritalStatus,ref = "Married")
#data1$Department<-relevel(data1$Department,ref = "Sales")

# MODEL 1 : multinomial logistic regression with the interaction term X1*
    X3
model<-multinom(BusinessTravel ~ MaritalStatus + Department +
    TotalWorkingYears +MaritalStatus:TotalWorkingYears ,family="binomial"
    ,data = data1)
summary(model)
```

```
17 OR<-exp(coef(model))
18
19 #statistical significance of coefficients
20 zvalues <- summary(model)$coefficients / summary(model)$standard.errors
21 pvalues<-pnorm(abs(zvalues), lower.tail=FALSE)*2
22
23 # Pseudo R-squared
24 PseudoR2(model, c("CoxSnell","McFadden", "Nagel"))
25
26 #effect
27 #plot(effect("MaritalStatus:TotalWorkingYears", model, xlevels=list(
       TotalWorkingYears=0:40)),
28 #multiline=TRUE, ylab="Probability(BusinessTravel)", rug=FALSE)
29
30 #summary statistics for the fitted model
31 Travel_Frequently <- cbind(summary(model)$coefficients[1, ],summary(model
       )$standard.errors[1, ],zvalues[1, ],pvalues[1, ],exp(coef(model)[1,])
       ,
32 exp(coef(model)[1,])-qnorm(0.975,0,1)*summary(model)$standard.errors[1,
       ],exp(coef(model)[1,])+qnorm(0.975,0,1)*summary(model)$standard.
       errors[1, ])
33 colnames(Travel_Frequently) <- c("Coefficient","Std. Errors","Z-stastic",
       "p-value","OR","lower bound","upper bound")
34
35 Travel_Rarely <- cbind(summary(model)$coefficients[2, ],summary(model)$
       standard.errors[2, ],zvalues[2, ],pvalues[2, ],exp(coef(model)[2,]),
36 exp(coef(model)[2,])-qnorm(0.975,0,1)*summary(model)$standard.errors[2,
       ],exp(coef(model)[2,])+qnorm(0.975,0,1)*summary(model)$standard.
       errors[2, ])
37 colnames(Travel_Rarely) <- c("Coefficient","Std. Errors","Z-stastic","p-
       value","OR","lower bound","upper bound")
38
39 kable(Travel_Frequently, "latex", booktabs = T) %>%
40 kable_styling(latex_options = c("striped", "scale_down"))
41 kable(Travel_Rarely, "latex", booktabs = T) %>%
42 kable_styling(latex_options = c("striped", "scale_down"))
43
44 # percentage of correct prediction
```

```
45
46 #Using sample_frac to create 70 - 30 slipt into test and train
47
48 #train <- sample_frac(data1, 0.7)
49 #sample_id <- as.numeric(rownames(train)) # rownames() returns character
       so as.numeric
50 #test <- data1[-sample_id,]
51
52 # Predicting the values for train dataset
53 #train$precticed <- predict(model, newdata = train, "class")
54
55 # Building classification table
56 #ctable <- table(train$BusinessTravel, train$precticed)
57
58 # Calculating accuracy - sum of diagonal elements divided by total obs
59 #round((sum(diag(ctable))/sum(ctable))*100,2)
60
61 # Predicting the values for train dataset
62 #test$precticed <- predict(model, newdata = test, "class")
63
64 # Building classification table
65 #ctable1 <- table(test$BusinessTravel, test$precticed)
66
67 # Calculating accuracy - sum of diagonal elements divided by total obs
68 #round((sum(diag(ctable1))/sum(ctable1))*100,2)
```

Βιβλιογραφία

[1] Agresti. A. (2007). An Introduction to Categorical Data Analysis, 2nd edition, *Wiley Series in Probability and Statistics, Wiley: New York.*

[2] Alboukadel K. (2017). Practical Guide To Cluster Analysis in R, 1st edition, *STHDA (Statistical Tools for High-throughput Data Analysis (http://www.sthda.com).*

[3] Collett D. (1991). Modelling Binary Data, 1st edition, *Chapman & Hall Statistics Text Series.*

[4] Fagerland, M., Hosmer D., Bofin, A. (2008). Multinomial goodness-of-fit tests for logistic regression models, *Statistics in Medicine*, 27, 4238–4253.

[5] Fagerland, M. W. (2009). Performance of Significance Tests, with Emphasis on Three Statistical Problems in Medical Research. Series of Dissertations Submitted to the Faculty of Medicine, No. 853, University of Oslo.

[6] Fagerland, M. W., & Hosmer, D. W. (2012). A generalized Hosmer–Lemeshow goodness-of-fit test for multinomial logistic regression models. *The Stata Journal*, 12(3), 447–453.

[7] Hallett D. C. (1999). Goodness of fit tests in logistic regression. Master's thesis, University of Toronto, Graduate Department of Community Health.

[8] Hamid H. A., Hassan A., Yap B. W., Amin N. A. M., (2018). Investigating The Power of Goodness-of-fit Test for Multinomial Logistic Regression Using K-means Clustering Technique, *AIP Conference Proceeding* 2013, 020004.

[9] Hamid H. A., Yap B. W., Xie X. J. and Ong S. H. (2017). Investigating the power of goodness of tests for multinomial logistic regression. *Communications in Statistics-Simulation and Computation*, 47, 4, 1039-1055.

[10] Hosmer Jr, D. W., Lemeshow, S., Sturdivant, R. X., (2013). Applied Logistic Regression, 3nd edition, *Wiley Series in Probability and Statistics. Wiley: New York.*

[11] Hussain N. J., Atheer J.A. (2015). Cluster Analysis as a Strategy of Grouping to Construct Goodness-of-Fit Tests when the Continuous Covariates Present in the Logistic Regression Model, *British Journal of Mathematics & Computer Science*, 10(1), 1-16.

[12] Peduzzi P., Concato J., Kemper E., Holford T., Feinstein A. (1996). A simulation study of the number of events per variable in logistic regression analysis, *Journal of Clinical Epidimiology*, 49(12), 1373-9.

[13] Pulkstenis E, Robinson TJ. (2002). Two goodness-of-fit tests for logistic regression models with continuous covariates. *Statistics in Medicine*, 21, 79-93.

[14] Pulkstenis, E., Robinson, T. J. (2004). Goodness-of-fit tests for ordinal response regression models, *Statistics in Medicine*, 23(6), 999–1014.

[15] Rodríguez, G. (2007). Lecture Notes on Generalized Linear Models, Logit Models for Binary Data. http://data.princeton.edu/wws509/notes/

[16] Tsiatis, A. A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika*, 67(1), 250–251.

[17] Wei Ma. (2018). A study of somes issues of goodness-of-fit tests for logistic regression, PhD diss., Florida State University, College of Arts and Sciences, Department of Statistics.

[18] Velden, Michel & Iodice D'Enza, Alfonso & Markos, Angelos. (2018). Distance-based clustering of mixed data, *WIREs Comput Stat*. 2019, 11:e1456.

[19] Xie X, Pendergast J, Clarke W. (2008). Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors, *Computational Statistics & Data Analysis*, 52(5), 2703–2713.

[20] Xie, X-J., Bian, A. (2009). Increasing the power: goodness-of-fit tests for ordinal response regression models with continuous covariates. *JP Journal of Biostatistics*, 3(3), 225–246.

[21] Ying Liu. (2007). On goodness-of-fit of logistic regression model. PhD diss., Kansas State University, College of Arts and Sciences, Department of Statistics.

[22] Κατσουδάκη Ε. (2016). Γραμμική και μη γραμμική παλινδρόμηση με εφαρμογές στην R, Μεταπτυχιακή διπλωματική εργασία, Πανεπιστήμιο Πατρών, Τμήμα Μαθηματικών.

[23] Μανούσου Κ. (2014). Στατιστική Ανάλυση Λοιμώξεων από το Πολυ-Ανθεκτικό Acinetobacter baumannii. Διπλωματική εργασία, Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών, Τομέας Μαθηματικών.

[24] Ντζούφρας Ι., Περπέρογλου Α. (2009). Εισαγωγή στην Βιοστατιστική και την Επιδημιολογία, Οικονομικό Πανεπιστήμιο Αθηνών & Πανεπιστήμιο Αιγαίου, Τμήμα Στατιστικής & Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοκοοικονομικών Μαθηματικών, Σημειώσεις.

[25] Ξενή Μ. (2016). Λογιστική Παλινδρόμηση & Διαχωριστική Ανάλυση, Μεταπτυχιακή διπλωματική εργασία, Πανεπιστήμιο Πατρών, Τμήμα Μαθηματικών.

[26] Στρατινάκης Ν. (2018). Εφαρμοσμένη Ανάλυση Συστάδων, Μεταπτυχιακή διπλωματική εργασία, Πολυτεχνείο Κρήτης, Σχολή Μηχανικών Παραγωγής και Διοίκησης.

[27] https://github.com/amzbst/Investigating-the-power-of-goodness-of-fit-tests-for-multinomial-logistic-regression/projects

[28] https://www.statisticssolutions.com/assumptions-of-logistic-regression/